Reasoning and knowledge - tradeoff between expressivity and efficiency

Ana Roxin

UNIVERSITÉ
BOURGOGNE FRANCHE-COMTÉ

SPIM

Raisonnement et connaissances
– à la recherche de l'équilibre entre expressivité et efficacité

Ana Roxin

Mémoire présenté afin d'obtenir le diplôme
d'**habilitation à diriger des recherches**

Université de Bourgogne Franche-Comté

Ecole doctorale SPIM (Sciences Pour l'Ingénieur et Microtechniques)

Discipline : Informatique, section CNU 27

Soutenance publique le 15 novembre 2018 devant un jury composé de :

Pr. Djamal BENSLIMANE	Université Claude Bernard Lyon 1	Rapporteur
Pr. Pascal MOLLI	Université de Nantes	Rapporteur
Pr. Michel PAINDAVOINE	Université de Bourgogne	Rapporteur
Pr. Frédérique LAFOREST	Telecom Saint-Etienne	Examinatrice
Pr. Florence SEDES	Université de Toulouse	Examinatrice
Pr. Bernd AMANN	Université Pierre et Marie Curie	Examinateur
Pr. Jean-Christophe LAPAYRE	Université de Franche-Comté	Examinateur
Pr. Christophe NICOLLE	Université de Bourgogne	Examinateur
Pr. Kokou YETONGNON	Université de Bourgogne	Membre invité

"If knowledge can create problems, it is not through ignorance that we can solve them"

Isaac Asimov

1 Préambule

Ce mémoire présente une partie des recherches que j'ai menées ou encadrées entre 2012 et 2018 au sein du laboratoire LE2I de l'Université de Bourgogne. Le mémoire est structuré autour d'une thématique centrale, à savoir l'ingénierie des connaissances, notamment comment représenter la connaissance d'un domaine de manière à ce qu'un ordinateur puisse l'interpréter et la manipuler de manière similaire qu'un humain le ferait. Les contributions présentées concernent la modélisation de savoir-faire métiers avec des approches formelles, décidables et en intégrant une sensibilité au contexte (de l'utilisateur ou d'utilisation). En relation avec ce verrou scientifique, de nombreuses questions sont encore ouvertes : comment modéliser de manière expressive et décidable une connaissance ? Comment identifier des connaissances vraies, avec de la valeur dans un environnement de traitement massif de données ? Comment dépasser les limites de l'hétérogénéité sémantique dans un environnement fortement distribué ?

Les recherches présentées ici ne concernent pas toutes des travaux achevés; l'idée avant tout est d'essayer de dégager le but commun vers lequel convergent mes recherches actuelles et à venir. Ce but, dont la spécification n'est pas un exercice facile, pourrait s'énoncer de la manière suivante: concevoir et implémenter des approches permettant de simuler un raisonnement humain sur des données réparties. Mes activités de recherche sont dans la continuité de mes travaux de thèse et portent sur la sensibilité au contexte, le profilage utilisateur, les systèmes de recommandation à base d'ontologies et l'interopérabilité sémantique des Systèmes d'Information (SI) au sens large.

1.1 Motivations

Qu'est-ce qui motive mon travail ? Depuis dix ans que j'exerce des activités de recherche, il m'a semblé légitime de m'interroger sur ce qui motive mes recherches, en dépit de la diversité et l'apparente divergence des directions de travail: raisonnement à partir de cas, modélisations sémantiques, coopération entre agents, données liées, etc. C'est en 2006, à la fin de mes études d'ingénieur informatique, que j'ai choisi un sujet de thèse CIFRE portant sur la sensibilité au contexte de l'utilisateur dans la recherche de services Web sémantiques. A cette époque, j'étais loin de m'imaginer la grande diversité et richesse des recherches à venir. Cette diversité se retrouve non seulement au niveau de mes axes de recherche, mais aussi au niveau des champs d'application couverts par mes recherches: systèmes de recommandation, systèmes d'information géographique (SIG), ou encore Building Information Modeling (BIM). Il est vrai que je m'intéresse à de nombreux champs sans autre motivation que la curiosité, scientifique ou intellectuelle. Toutefois, l'ensemble des thématiques de recherche abordées adresse la question centrale de comment transmettre à un ordinateur des connaissances humaines, tout en lui permettant de raisonner sur ces connaissances, comme le ferait un utilisateur humain. Il s'agit donc d'étudier les approches existantes et d'en proposer des nouvelles concernant la spécification, l'intégration et l'exploitation de connaissances métiers dans des environnements intelligents. L'ingénierie des connaissances (à travers des technologies sémantiques et des données liées) se retrouve donc au centre de mes recherches.

Une première interrogation apparaît au travers de ces lignes: qu'est-ce qu'on entend par "connaissances" et comment elles se différencient des données ou des informations ? Il existe en effet une distinction claire entre les concepts de "données", "informations", "connaissances" et "sagesse", présentée dans (Zack 1999), et qui reprend la hiérarchisation de ces concepts faite par Ackoff en 1989 (Ackoff 1989), à savoir DIKW pour *Data, Information, Knowledge, Wisdom*. Selon cette définition, les données sont des observations (ou des "éléments obtenus à partir de mesures de variables" selon (ISO772 2011)) et les informations se définissent en tant que "données placées dans un contexte" (Kemp 1999). Définir ce qu'est la connaissance n'est pas aisé; toujours d'après l'ISO les "connaissances" représentent des "faits, informations, vérités, principes ou compréhensions acquis grâce à l'expérience ou à l'instruction" (ISO17027 2014). En d'autres termes, les connaissances sont des informations augmentées d'expériences, de réflexions ou autres pratiques (Ericksons S. 2014). Par rapport à ces définitions, la "sagesse" est souvent définie en tant qu'intelligence de niveau supérieur, allant au-delà des connaissances basiques. L'ingénierie des connaissances correspond au domaine étudiant les approches de transformation des données et des informations en connaissances, qu'elles soient basées sur le l'apprentissage (automatique ou supervisé) ou sur des approches sémantiques. L'ingénierie des connaissances étend la "simple» gestion de connaissances; la difficulté réside dans le fait que les systèmes d'information d'aujourd'hui échangent facilement des données et des informations explicites.

Or, les connaissances sont souvent enfouies dans des informations implicites, et arriver à les représenter de manière à ce qu'une machine puisse les interpréter de manière similaire à un humain n'est pas trivial.

Représenter la connaissance d'un domaine de manière à ce qu'un ordinateur puisse l'interpréter et la manipuler revient à créer un modèle de cette connaissance, c'est-à-dire une spécification structurée de faits avérés (ou connaissances vraies) pour le domaine considéré. Ledit modèle de connaissance est par définition incomplet et de natures variées. Un tel modèle est construit par un modélisateur, selon une problématique donnée, avec un but bien défini. Les choix de modélisation seront impactés par le problème visé: le modélisateur peut être amené à ne spécifier que les connaissances nécessaires au problème, notamment pour optimiser l'algorithme sous-jacent. La modélisation de connaissances d'un domaine représente une activité humaine, hautement impactée par les connaissances du modélisateur dans le domaine visé, mais aussi par ses préférences en termes de modélisation. Dès lors, pour un domaine donné, il ne peut y avoir un modèle unique de la connaissance; en général, il y aura autant de modèles que de modélisateurs. Le problème d'hétérogénéité que ces modèles amènent apparaît clairement: comment savoir si deux représentations parlent de la même "chose" si elles définissent la "chose" de la même manière ou si elles présentent des points de vue concurrents ou alors approximatifs ? Mes recherches s'efforcent de répondre à cette problématique en utilisant les technologies dites du "Web sémantique" et l'approche "données liées". Basées sur une pile de standards poussés par le W3C, les approches sémantiques permettent de composer des ontologies ou des spécifications explicites et formelles des connaissances d'un domaine donné (ceci sera expliqué dans la Section 3.1). Malheureusement, la conception d'ontologies seule ne permet d'atteindre l'interopérabilité sémantique.

Dès lors, une deuxième interrogation se profile: qu'entend-t-on par "*interopérabilité*" et plus particulièrement qu'est-ce que ***l'interopérabilité sémantique*** ? L'ISO définit l'interopérabilité comme la "capacité que possèdent des systèmes à fournir des services et à recevoir des services d'autres systèmes et à utiliser les services ainsi échangés afin de fonctionner ensemble de manière efficace" (ISO17261 2012). D'un point de vue machine, implémenter une interopérabilité revient à relier deux systèmes informatiques hétérogènes, afin qu'ils puissent collaborer, ce qui implique un accès réciproque à leurs ressources. L'interopérabilité sémantique peut dès lors être définie comme la capacité d'un système informatique ou d'une machine d'interpréter des ressources d'un autre système ou d'une autre machine, et ce de la même manière. L'interopérabilité sémantique revient en quelque sorte à implémenter un échange de connaissances sans coutures, entre deux systèmes informatiques.

C'est dans ce contexte que se situent les travaux de recherche présentés dans ce manuscrit. Par rapport à cette problématique centrale, je me suis plus particulièrement intéressée à la ***fédération de connaissances réparties par le biais d'ontologies***. Les ontologies, en tant que spécification formelle et explicite d'une conceptualisation commune d'un domaine de connaissances, fournissent un niveau d'expressivité permettant de s'affranchir des problèmes liés à l'hétérogénéité syntaxique et schématique. Toutefois les ontologies seules ne permettent pas d'atteindre une interopérabilité sémantique. Combinées avec l'application des principes des données liées, les ontologies deviennent des graphes sémantiques liés entre eux créant ainsi un continuum d'expressivité. Porté à l'échelle du Web, cela permettrait la création d'un graphe global géant que des agents informatiques pourraient "explorer" afin de découvrir des connaissances implicites.

Au travers de mes recherches, j'ai accordé une attention particulière à l'impact que les choix faits au niveau de la représentation des connaissances pouvaient avoir sur l'efficacité de l'implémentation. En effet, sans contraintes d'implémentation, sans objectifs clairs à atteindre, en termes d'interopérabilité sémantique notamment, le modélisateur dispose d'une liberté totale dans le formalisme de représentation suivi. Cela peut s'avérer intéressant d'un point de vue purement recherche, mais concevoir des représentations sans se soucier de leur utilisation future, n'est pas le choix que j'ai fait. Dès lors, dans chacune des recherches menées depuis ma thèse, j'ai toujours essayé d'atteindre un équilibre entre l'expressivité des modèles conçus et l'efficacité d'implémentation. Comme présenté dans le chapitre décrivant les recherches menées durant le co-encadrement des travaux de Madame HOPPE (voir chapitre 3.2), une trop grande expressivité du modèle de représentation des connaissances n'en facilite pas l'implémentation. Nous verrons que la modularité d'un tel formalisme de représentation est importante lorsqu'il s'agit de l'intégrer à d'autres formalismes, mais aussi qu'en termes d'interopérabilité sémantique, la fédération de tels formalismes offre plus d'avantages.

1.2 Plan du document

Ce manuscrit est une présentation de mes travaux de recherche menés depuis mon doctorat, visant à mettre en évidence mon expérience dans l'animation d'activités de recherche non individuelles. Elle se scinde en deux parties :

1. La première partie résume mes activités d'enseignant-chercheur depuis l'obtention de mon doctorat en 2009, plus particulièrement depuis ma nomination en tant que maître de conférences en informatique à l'Université de Bourgogne, au sein du département IEM (Informatique Électronique et Mécanique) et du laboratoire LE2I[1] (Électronique, Informatique, Image). Cette partie se compose d'un curriculum vitae synthétique, d'un bilan des activités d'enseignement, d'un bilan scientifique et d'un bilan des activités administratives. Le bilan scientifique comprend la liste de mes publications scientifiques, la description de mes implications dans les directions de thèses de doctorat, et la liste des projets de recherche nationaux et européens auquel j'ai pris part en tant que participant ou responsable. Le bilan des activités administratives comprend les responsabilités pédagogiques (dont la responsabilité de l'option "Mobile" de la licence LPMI), et les responsabilités scientifiques (dont l'implication dans des organismes de standardisation, l'expertise de projets de recherche internationaux et la relecture d'articles de revues et de conférences). Plusieurs sections sont consacrées à la présentation des apports de la recherche au travers d'activités de publication, transfert industriel, diffusion des connaissances, gestion de projets scientifiques et notamment encadrement doctoral (thèses de doctorat) et scientifique (master recherche, master 2, etc.).

2. La deuxième partie retrace l'évolution de mes activités de recherche, et présente le socle sur lequel sont construits mes projets de recherche prospectifs. Cette seconde partie cherche à donner une vision d'ensemble des recherches menées depuis mon doctorat, sous la forme d'une discussion sur l'***approximation entre expressivité et efficacité des modèles et approches sémantiques***. Ceci représente le cœur de ma problématique recherche, à savoir la ***modélisation de savoir-faire métiers avec des approches formelles et décidables, de même que l'intégration et l'exploitation de ces connaissances métiers pour le développement d'applications sensibles au contexte*** (utilisateur comme utilisation). Cette partie comprend trois principaux chapitres (tels que décrits ci-dessous), chacun d'eux présentant les recherches menées par rapport aux axes considérés :

    a. La section 3.1 présente le cadre de travail s'appliquant aux travaux présentés dans ce manuscrit. J'y présente les concepts nécessaires à la compréhension du Web sémantique : vision du Web de données, langages de description d'ontologies, langages de description de règles, etc. Un résumé des principales caractéristiques des données sémantiques par rapport aux données issues d'approches relationnelles ou orientées-objet clôturera cette section.

    b. La section 3.2 présente les recherches effectuées dans le cadre du projet MindMinings (co-encadrement de la thèse de doctorat de Madame HOPPE, de 2012 à 2015). Ces travaux illustrent l'impact qu'ont les choix de modélisation et structuration des connaissances sur l'implémentation. Les recherches comprises par ce volet étudient les manières possibles de modéliser des connaissances afin d'être le plus fidèle à l'expertise métier devant être transmise à l'ordinateur (axe 1). La question recherche sous-jacente est: ***comment modéliser de manière expressive et décidable une connaissance tout en réduisant la distance d'apprentissage et en s'adaptant à un contexte ?***. J'y présente notre approche pour modéliser l'incertitude dans une ontologie et son application dans un système de profilage à base de règles logiques.

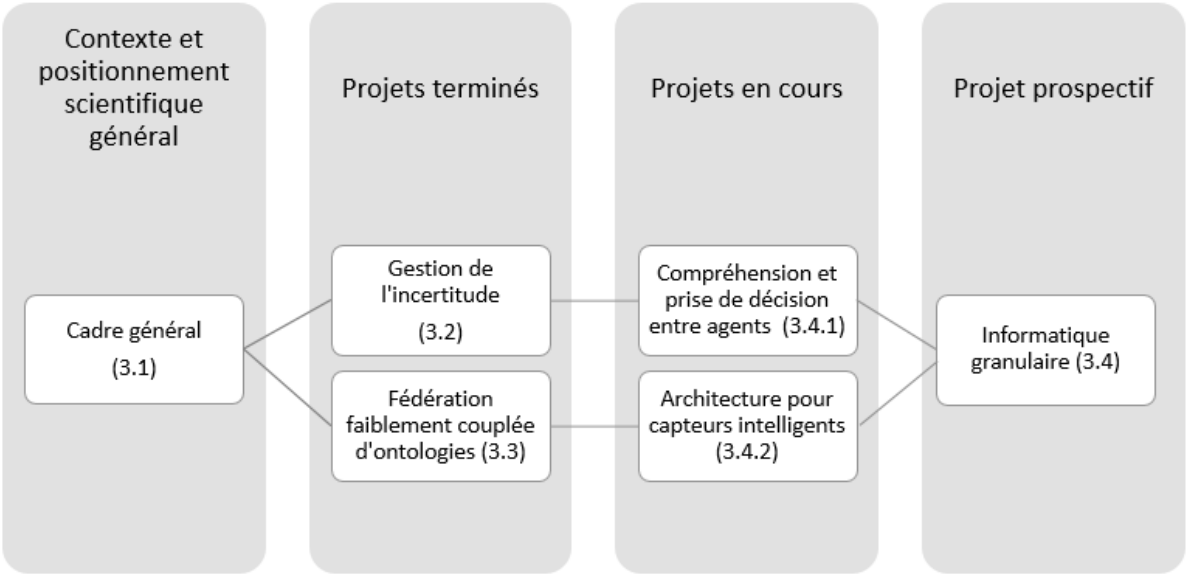
    c. La section 3.3 présente les recherches menées durant le co-encadrement de la thèse de Monsieur MENDES de FARIAS. J'y étudie les contraintes soulevées par un haut niveau d'efficacité de l'implémentation sur la structuration et la représentation des connaissances. Les recherches menées dans ce contexte visent à augmenter l'efficacité des approches à base d'ontologies, tout en réduisant la complexité ou encore le temps d'exécution des requêtes. La question de recherche sous-jacente vise à ***dépasser les limites de l'hétérogénéité sémantique dans un environnement réparti*** (axe 2). J'y présente notre approche pour la

[1] UMR CNRS 6306 (jusqu'en 2017), puis FRE2005 en 2018

fédération faiblement couplée d'ontologies, ainsi que notre approche pour l'amélioration du temps d'exécution des requêtes adressées dans le contexte d'une telle fédération.

d. La section 3.4 décrit mon projet de recherche prospectif, construit sur la base des recherches menées et en cours. J'y présente les verrous et approches investigués dans le contexte de deux thèses de doctorat que je co-encadre actuellement (thèses en cours), visant la mise en place, d'une part, d'un système multi-agents pour l'évaluation de scénarii de gestion de catastrophe (travaux de Madame PRUDHOMME, débutés en 2016) et, d'autre part, d'une architecture de capteurs sémantiques dans un environnement bâti (travaux de Monsieur ARSLAN, débutés en 2017). Les travaux de recherche associés trouvent des connexions avec les deux premiers axes exposés. En effet, le but visé ici est ***faire interopérer différents modèles et données*** (axe 2) afin de ***délivrer des réponses pertinentes par rapport à un contexte donné*** (axe 1). Sur la base des perspectives et verrous restants dans les travaux menés précédemment, je dégage une orientation future concernant la ***définition de niveaux de granularité lors de la conception d'ontologies***. Ces deux projets illustrent le besoin d'approches nouvelles, notamment besoin d'approches granulaires.

Ainsi, la figure suivante peut être utilisée comme grille de lecture pour ce manuscrit.



Pour chacun de ces chapitres, le plan suivant sera observé:

- Sommaire
- Perspective historique
- Description du domaine et des problèmes étudiés
- Rappel de l'état de l'art spécifique à la thématique considérée
- Approche et résultats
- Conclusion et ouvertures

Les différentes équations présentant des énoncés et contraintes en logique de description utilisent le modèle de syntaxe présenté dans (Krötzsch et al. 2012).

1.3 Remerciements

Cette Habilitation à Diriger des Recherches marque une étape symbolique dans ma carrière scientifique, nécessaire pour faire le bilan des années passées et indispensable pour préparer au mieux les années à venir. Ces travaux sont le fruit de nombreuses nuits de travail solitaire, mais également d'encore plus nombreuses rencontres, discussions et collaborations que j'espère avoir réussi à synthétiser au sein de ce mémoire de manière à lui donner la richesse et la profondeur souhaitées. Je saisis donc cette opportunité pour remercier vivement l'ensemble des personnes qui ont contribué à l'élaboration de ces réflexions.

Je voudrais remercier, en premier lieu, les étudiants que j'ai encadrés. C'est grâce à eux que j'ai appris la direction de la recherche et c'est en partie grâce à eux que je peux aujourd'hui présenter ce mémoire. Merci donc, dans l'ordre d'apparition, à Damien Zomahoun, Anett Hoppe, Jordan Simonot, Eric Guillon, Fabien Chaintreuil, Ali Almoukahal, Jiras Biokou, Tarcisio Mendes de Farias, Youssef Keryakos, Claire Prudhomme, Muhammad Arslan, Antoine Candy, Antonin Annet, Jimmy Grossiord, Vincent Imbeaux, Ali Osman Soykurt, Benoît Barnoux, Magali Barrue-Bassin.

Si j'ai beaucoup appris des étudiants, j'ai encore plus appris de mes collègues. Je tiens à remercier deux personnes qui m'ont particulièrement accompagné ces dernières années, qui ont su m'aider à affiner ma contribution scientifique, à prendre le recul nécessaire et à faire face aux difficultés de la recherche. Je tiens à remercier en premier lieu Christophe Nicolle, Professeur à l'Université de Bourgogne, qui m'a accordé toute sa confiance pour le co-encadrement de deux thèses de doctorat, tout en me montrant ce qu'était la liberté d'un chercheur. L'autonomie et la prise de responsabilités qu'il m'a accordées m'ont permis de progresser. Ma gratitude va aussi à Kokou Yetongnon, Professeur à l'Université de Bourgogne, dont les conseils prodigués m'ont évité bien des écueils. Mes remerciements vont aussi à Dominique Ginhac, Professeur à l'ESIREM et directeur du laboratoire LE2I, pour m'avoir aidé dans les différentes démarches administratives et pour la confiance qu'il m'a témoignée.

J'adresse mes vifs remerciements à Djamal Benslimane, Professeur à l'Université Claude Bernard Lyon 1, Pascal Molli, Professeur à l'Université de Nantes, ainsi qu'à Michel Paindavoine, Professeur à l'Université de Bourgogne, qui m'ont fait l'honneur d'être rapporteurs sur ce manuscrit, en sacrifiant une partie de leur temps pour se pencher avec indulgence sur ce manuscrit. Leurs remarques pertinentes ont contribué à son enrichissement. Tous mes remerciements vont également à Frédérique Laforest, Professeur à Telecom Saint-Etienne, Florence Sèdes, Professeur à l'Université de Toulouse, Berdt Amann, Professeur à la Sorbonne, et Jean-Christophe Lapayre, Professeur à l'Université de Franche-Comté, pour avoir accepté de participer à ce jury et de m'avoir permis d'approfondir mes réflexions sur ce sujet de recherche.

L'Université de Bourgogne et le département IEM (Informatique, Électronique, Mécanique) ont été un lieu d'émulation collective et de partage qui ont particulièrement contribué à la richesse professionnelle et humaine de ces dernières années. Je souhaite remercier vivement Florence Mendes pour ses corrections précieuses lors de la rédaction de ce mémoire, ainsi que mes collègues de l'aile H, Ouassila Labbani-Narsis, Denis Pellion, Dominique Arnoult ainsi que l'ensemble des membres du futur Laboratoire Informatique de Bourgogne (LIB).

J'aimerais exprimer mes remerciements aux différents personnels administratifs sans l'aide desquels les travaux mentionnés dans ce manuscrit ne se seraient pas déroulés dans d'aussi bonnes conditions. Je pense notamment à Mélanie Arnoult et Dounia Radi, secrétaires du laboratoire LE2I, qui ont géré les ordres de missions, les remboursements et les budgets projets. Je pense aussi à Claire Da Silva Moreira, qui est l'unique aide administrative dont je dispose pour le projet ANR McBIM, et sans laquelle je n'aurais pas pu gérer le projet.

Enfin, je souhaite créditer les différents institutions et entreprises qui ont financièrement soutenu le bon déroulement de nos travaux : la société Ezakus, la société Active3d, l'association mediaConstruct, le CSTB et enfin mon employeur, l'Université de Bourgogne, ainsi que la communauté d'universités qu'elle a rejoint depuis 2015, à savoir la COMUE UBFC (Univ. Bourgogne Franche-Comté), en particulier son Président, Monsieur Nicolas Chaillet, ainsi que son Premier Vice-Président, Monsieur Olivier Prévot.

Ana Roxin
ana-maria.roxin@ubfc.fr

Sommaire

2 Activités d'enseignant-chercheur menées depuis le doctorat

2.1 CV synthétique

2.1.1 Identification

NOM Prénom	ROXIN Ana-Maria
Fonction actuelle	Maître de conférences en Informatique, 5e échelon Titulaire de la PEDR depuis Décembre 2017
Qualifications CNU (2009)	27e section : 10227203592 61e section : 10261203592
Composante de rattachement	UFR Sciences et Techniques, Université de Bourgogne
Laboratoire de rattachement	Laboratoire LE2I (Electronique, Informatique et Image) FRE2005 Université de Bourgogne, UBFC

2.1.2 Parcours professionnel

Période	Fonction	Employeur
Depuis Septembre 2012	Maître de conférences en informatique	Université de Bourogne, UFR Sciences et Techniques, 9 allée Alain Savary, 21000 DIJON
Novembre 2010 – Juin 2012	Chercheur postdoc	UFR-STGI, UMR CNRS 6249, 4 Place Tharradin, 25200 MONTBELIARD
Novembre 2009 – Août 2010	ATER (Attachée Temporaire d'Enseignement et de Recherche)	UTBM, Laboratoire Systèmes et Transports (SeT) – EA 3317, Technopôle, 90010 BELFORT
Mai 2009 – Octobre 2009	Vacataire	UTBM, Laboratoire SeT – EA 3317, Technopôle, 90010 BELFORT
Octobre 2006 – Avril 2009	Doctorant CIFRE / Chef de projet (CDD)	Pimentic sas – 6, av. des Usines, 90000 BELFORT
Février 2006 – Août 2006	Stagiaire ingénieur	Orange Labs – Equipe RESA/NET – 90000 BELFORT

2.1.3 Formations universitaires

Diplôme de doctorat "Protocole de découverte, sensible au contexte, pour les services Web sémantiques"		
Date	30 Novembre 2009	
Etablissement	Université de Technologie de Belfort-Montbéliard (UTBM)	
Mention	Très honorable	
Directeur de thèse	Maxime WACK – MCF HDR	UTBM – 61e section
Co-directeur de thèse	Jaafar GABER – MCF	UTBM – 27e section
Rapporteurs	Daniel JOLLY - PR	Université d'Artois – 61e section
	Jean-Marie PINON – PR	Institut National des Sciences Appliquées (INSA) de Lyon – 27e section
Examinateurs	Alexandre CAMINADA – PR	UTBM – 27e section
	Jacques SAVOY – PR	Université de Neuchâtel (UNINE) – 27e section

Diplôme d'ingénieur informatique, spécialité "Réseaux Mobiles et Systèmes Embarqués"	
Date	Juin 2006
Etablissement	Université de Technologie de Belfort-Montbéliard (UTBM)
Filière	Réseaux Mobiles et Systèmes Embarqués

2.1.4 Formations continues

Date(s) / Durée	Titre	Intervenant(s)
Mardi 18 octobre 2016, à Lyon ~4h	Formation à l'écriture de projets Horizon 2020 (How to Write a Competitive Proposal for Horizon 2020)	Sean McCarthy, société Hyperion Ltd.
3 modules, sur 2 jours chacun : 15-16 mars / 24-25 mai / 6-7 juillet 2016 ~42h	La professionnalisation pédagogique des directeurs de thèse: de la connaissance de l'étudiant et de l'analyse du processus de thèse à l'appropriation des outils d'accompagnement	Philippe BELPAEME, Formateur Société Belpaeme Conseil Hélène BARBIER-BRYGOO, Directrice de recherche CNRS, Directrice de l'Institut des Sciences du Végétal
Vendredi 31 mai 2013 7h	Triche et plagiat des étudiants : définir, détecter, prévenir	Mallory SCHAUB, conseillère pédagogique, université de Genève, Suisse
Vendredi 3 mai 2013 7h	La supervision doctorale : Comment améliorer ma pratique d'encadrement des doctorants	Denis BERTHIAUME, Consultant en Développement de l'Enseignement Supérieur

2.1.5 Compétences techniques et connaissances linguistiques

Langages informatiques – programmation et modélisation	
Orientés-objet	Java, Java pour Android, C++
Web	JavaScript, HTML 4/5, XML, CSS, PHP, ASP, .NET, Responsive Web Design, node.js, Apache Cordova, Xamarin, ionic
Systèmes d'informations	UML, MERISE/MERISE 2, SQL, Oracle
Autres langages	C, LISP, Prolog, ActionScript, Perl, Python, RDFS, OWL, OWL-S, SWRL, N3
Logiciels informatiques	
Bureautiques	Suites Office & OpenOffice, Mindjet MindManager
Web	Drupal, SPIP, Adobe Flash, Adobe Dreamweaver
Bases de données	WinDesign, Access
Traitement d'image	Gimp, Adobe Photoshop, Adobe Illustrator, Adobe Fireworks
Mathématiques	Matlab 7, Maple 5
Systèmes d'Information Géographique (SIG)	MapInfo 7-9, MapPoint, ArcGIS
eLearning	Learning Content Management System (LCMS) e.g. Moodle, Adobe Connect Pro
Connaissances linguistiques	
Anglais	Bilingue, obtention du TOEIC avec un score de 955
Roumain	Bilingue, langue maternelle
Espagnol	Bon niveau général (lu, parlé, écrit), obtention du DELE inicial
Allemand	Niveau intermédiaire (lu, parlé, écrit)

Liens additionnels	
Profil LinkedIn	https://www.linkedin.com/in/ana-roxin-6ba862a/
Profil ORCID	https://orcid.org/0000-0001-9841-0494
Profil Google Scholar	https://scholar.google.fr/citations?user=3Q0eunwAAAAJ&hl=fr
Profil ResearchGate	https://www.researchgate.net/profile/Ana_Roxin
Liste complète des publications	https://cv.archives-ouvertes.fr/ana-roxin

2.2 Activités d'enseignement

2.2.1 Synthèse des activités d'enseignement

Niv.	Formation	Cours
L	Licence PRO MI (Métiers de Internet)	**Cours "Programmation Android"** (Responsable UE) > Présentation de la plateforme Android (historique, terminaux, noyau Linux, fonctionnalités de base, machine virtuelle, etc.) > Présentation des concepts de base (vues, contrôles, activités et sous-activités, fichier de configuration, etc.) > Composants de base des applications (services, fournisseurs de contenus, widgets, objets Intent, récepteurs d'intentions, filtres d'intention, notifications, alarmes, etc.) > Les ressources Android > Les menus et les menus contextuels > Interactions avec l'utilisateur (gestion du clic sur un bouton, afficher et enrichir des notifications, gestion de la vibration du terminal, appel d'activités système, etc.) > Création d'applications avec IHM (concept d'interface, les layout ou gabarits, les vues, les ListView, les boîtes de dialogue, etc.)
		Cours "Responsive Web Design" (sites pour terminaux mobiles) > Présentation du contexte mobile et des contraintes connexes > Approches d'adaptation existantes > Conception et codage d'interfaces en utilisant HTML5 et CSS3 > Responsive Web Design > Requêtes média CSS (CSS media queries) > Définir des mises en page fluides > Construire des sites Web spécifiques pour les terminaux mobiles > Détecter les capacités du client > Utiliser le stockage local du terminal > Positionner le mobile > Utiliser les URI de données > Détection de l'User Agent
	3e année ESIREM[2] (école d' ingénieurs)	Introduction au développement Web (HTML, CSS, Responsive Web design, JavaScript pour mobiles)
	L1	**HTML5, JavaScript, Perl** > Expressions régulières en Perl > Introduction aux expressions régulières > Opérateurs de base (affectation, comparaison, substitution, recherche) > Construire des expressions régulières (intervalles, sélection avec alternative, multiplicateurs de sélecteurs, mémoriser une sélection, etc.)
M	M2 Pro BDIA (Bases de Données et Intelligence Artificielle)	**Cours "Web de données liées"** > Le déluge de données et la justification des données liées dans ce contexte > Les principes des données liées > Nommer des éléments avec des URI > Rendre les URI déréférençables (URI 303 et URI avec ancre) > Le modèle de données RDF > Les formats de sérialisation RDF (RDF/XML, N-Triples, Turtle, Notation3 ; NQuads)

[2] Ecole d'ingénieurs en Matériaux, Développement Durable et Infotronique

		> Les différents types de liens vers d'autres éléments (liens de relation, d'identité et de vocabulaire) > Le Web de données : démarrage et topologie > Considérations de conception sur les données liées > Utiliser des URI comme des noms pour les éléments > Forger des URI HTTP > Lignes directrices pour créer de bonnes URI > Comment décrire les éléments avec RDF > Triplets littéraux et liens sortants > Liens entrants > Triplets décrivant les ressources liées > Triplets sur la description > Décrire un jeu de données liées > Métadonnées de provenance > Licences, renonciations et normes pour les données > Choisir et utiliser les vocabulaires pour décrire des données > Publier un jeu de données liées > Fournir des données liées sous la forme de fichiers RDF/XML statiques > Fournir des données liées comme du RDF intégré dans des fichiers HTML > Fournir du RDF et du HTML avec des scripts serveur ad hoc > Fournir des données liées à partir de bases de données relationnelles > Fournir des données liées à partir de stockage de triplets RDF > Fournir des données liées en adaptant des applications ou des API web > Tester et déboguer les données liées > Liste de contrôles pour la publication de données liées
		Cours "Cloud computing et Sécurité" > Emergence du cloud computing > Concepts du cloud computing > Terminologies : SaaS, PaaS, IaaS, etc. > Modèle du cloud par rapport au modèle classique du software > Concepts de cloud privé, hybride, public, etc. > Usages (mashups, SaaS, IaaS et PaaS) > Remarques sur le cloud du point de vue des utilisateurs, des décideurs, des informaticiens, etc. > Exemples de déploiements (gestion de comptes d'utilisateurs, fédération d'identité, intégration de clouds, structuration des SI dans le cloud, héberger des applications critiques dans le cloud, etc.) > Présentation et caractérisation des principales offres actuelles dans le cloud computing (positionnement des grands acteurs IT, collaboration unifiée, services de FrontOffice et de BackOffice, etc.) > Sécurité des systèmes d'information : authentification (Kerberos), chiffrage, bonnes pratiques (ANSSI), réglementation (GDPR), utilisation de services cloud pour se protéger, étude des nuages noirs (e.g. botnet).
	M1 Multimédia	**HTML5/CSS3 et JavaScript pour le multimédia, Streaming audio, vidéo, Perl, indexation textuelle, algorithme PageRank** > Streaming audio et vidéo > Introduction et histoire du streaming > Techniques de streaming (mise en mémoire tampon, adaptation dynamique aux variations de débit, diffusion unicast et multicast, pseudo streaming, > Architectures et formats (Windows Media, RealMedia, QuickTime, MPEG-4)

		> Protocoles de streaming (RTP Real Time Protocol, RTCP Real Time Control Protocol, RVSP Ressource reSerVation Protocol, RTSP Real Time Streaming Protocol) > Recherche et indexation d'informations > Méthodes d'indexation (manuelle ou automatique) > Modèles pour la recherche d'informations (modèle booléen, modèle vectoriel, modèle probabiliste) > Adaptation de sites Web pour terminaux mobiles : Responsive Web Design, jQuery, HTML5/CSS3 avancé
ED SPIM[3]	Module **"Web of Data"**	Principes des données liées, vocabulaires pour l'annotation de données, publier et consommer des données liées
	Module **"Cloud Computing and Big Data"**	Modèles du cloud, architectures cloud, application dans les entreprises

Année	CM	TD	TP	CI	Total heures
2013/2014	23,5%	32,5%	21%	23%	192
2014/2015	26,5%	40,5%	15%	18%	232
2015/2016	12%	21,7%	29,5%	36,9%	287
2016/2017	16,6%	20,5%	27,9%	34,9%	285
2017/2018	15,4%	49,1%	35,5%	0%	256

2.2.2 Responsabilités pédagogiques

> Pour la licence Pro MI (LPMI)
  > Intégration dans l'équipe pédagogique - participation aux réunions pédagogiques, jurys fin d'années, conseils de perfectionnement, soutenances de stages, etc.
  > En 2014, j'ai proposé la création d'une option "Développement pour terminaux mobiles" pour les étudiants de la LPMI. Cette option a été validée par le CA de l'Université de Bourgogne en 2015, et a ouvert à la rentrée de septembre 2015. Je suis responsable de cette option "Mobile".
  > A partir de 2017, je suis responsable du module UE9 "Programmation Android"

> Pour le Master 2 Pro BDIA
  > Intégration dans l'équipe pédagogique - participation aux réunions pédagogiques, jurys fin d'années, conseils de perfectionnement, etc.

> J'encadre tous les ans plusieurs projets étudiants, dans le cadre des différentes formations : projets tuteurés et contrats de professionnalisation pour la Licence Pro MI et le Master 2 BDIA ; stages d'étudiants en Master 2 ; élèves ingénieurs en stage de 4e et 5e année à l'ESIREM ; étudiants en VAE (Validation des Acquis de l'Expérience) en collaboration avec l'ESIREM.

> Participation aux réunions du département IEM (Informatique, Eléctronique et Mécanique)

> Participation aux Assemblées Générales du laboratoire LE2I depuis 2013

2.2.3 Activités et responsabilités administratives

> J'ai pris part à la conception du dossier PARI STIC/SANTE 2012 N°9 pour la partie demande de FABER "équipement et contrat d'étude"

> Participation aux réunions d'équipe

[3] Ecole Doctorale SPIM (Sciences Pour l'Ingénieur et Microtechniques)

2.3 Activités scientifiques

2.3.1 Encadrements

2.3.1.1 Encadrement doctoral

2.3.1.1.1 Vue synthétique de l'encadrement doctoral

Depuis ma nomination aux fonctions de Maître de Conférences en 2012, j'ai participé au co-encadrement de 5 thèses de doctorat, dont 3 ont déjà été soutenues et 2 sont actuellement en cours.

Thèses de doctorat co-encadrées				
Index	**Etudiant(e) encadré(e)**	**Directeur(s) de thèse**	**Dates (Début - Fin)**	**Résumé**
1	Damien ZOMAHOUN	Kokou YETONGNON	17/11/2011 – 05/06/2015	Développement d'un système d'annotation et de recommandation pour les images, en utilisant les technologies sémantiques
2	Anett HOPPE	Christophe NICOLLE	01/10/2012 – 25/03/2015	Profilage sémantique et dynamique d'internautes en exploitant des techniques d'apprentissage d'ontologies (ontology learning)
3	Tarcisio MENDES de FARIAS	Christophe NICOLLE	01/10/2013 – 31/10/2016	Gestion sémantique et modélisation contextuelle du BIM Active3D en situation de mobilité (NomadBIM)
4	Claire PRUDHOMME	Christophe CRUZ	20/10/2015 – (en cours)	Conception d'un SIG sémantique, puis de son implémentation dans un contexte de gestion de crise
5	Muhammad ARSLAN	Christophe CRUZ / Dominique GINHAC	13/02/2017 – (en cours)	Architecture Sémantique de Capteurs intelligents Adaptatifs en Environnement bâti

2.3.1.1.2 Vue analytique de l'encadrement doctoral

Le tableau ci-dessous présente une vue détaillée des thèses que j'ai co-encadrées depuis 2012. Pour chacune d'entre elles, je mentionne le pourcentage de mon encadrement, le type de financement ainsi que le devenir du doctorant. La colonne "Lien avec le projet scientifique" discute la correspondance entre le sujet de thèse co-encadré et mes orientations de recherche définies dans mon projet de recherche.

Informations générales		**Lien avec le projet scientifique**
NOM Prénom	**ZOMAHOUN Damien**	La contribution de cette thèse est focalisée sur l'annotation collaborative d'images avec des techniques sémantique, et l'exploitation de ces annotations pour la recommandation d'images. La représentation sémantique repose sur une ontologie d'application dérivée d'une ontologie générique. L'annotation sémantique collaborative que nous proposons consiste à faire émerger la sémantique des images à partir des sémantiques proposées par une communauté d'annotateurs. L'application ainsi conçue répond à un besoin spécifique d'une entreprise spécialisée dans l'annotation manuelle d'images à base de mots-clés. L'approche sémantique proposée dans le cadre de cette thèse a permis de mettre en évidence les avantages des technologies sémantiques par rapport aux approches basées sur des mots-clés. Les travaux en lien avec cette thèse s'inscrivent dans l'axe "Sensibilité au contexte de l'utilisateur" de mes
Titre de la thèse (FR)	Emergsem : une approche d'annotation collaborative et de recherche d'images basée sur les sémantiques émergentes (Lien HAL)	
Date Début – Date Fin	17/11/2011 – 05/06/2015	
% d'encadrement	50%	
Nom et % d'encadrement du directeur	YETONGNON Kokou 50%	
Type de financement	Entreprise gabonaise	

Devenir du doctorant	ATER à l'Université de Bourgogne 2016-2017. A rejoint par la suite l'entreprise ayant financé sa thèse de doctorat.	orientations de recherche (puisqu'on a défini un système de recommandation d'images). Pour les mener à bien, j'ai utilisé des compétences acquises suite à mes recherches sur les technologies sémantiques pour l'eLearning et pour le raisonnement à partir de cas.
NOM Prénom	**HOPPE Anett**	Cette thèse adresse le thème de la sensibilité au contexte et du profilage utilisateur à travers des modèles ontologiques. Le profil utilisateur est construit à partir du traitement automatique de données de navigation. Une fois ce profil construit, il est apparié à des segments marketing par le biais de règles logiques. Un profil utilisateur peut être lié à différents segments marketing, chaque lien ayant une pondération différente. La principale contribution de cette thèse se situe dans la gestion de l'incertitude avec des technologies sémantiques. Alors que les langages d'ontologies permettant de gérer des notions probabilistes (e.g. énoncés qui ne sont ni vrai ni faux, mais qui ont un pourcentage de probabilité d'être "vrais") sont très peu utilisés en pratique (e.g. pr-OWL[4]), notre approche emploie les chaînes de propriétés pouvant être définies en OWL 2 DL afin d'associer des pondérations à des énoncés. Cette approche permet de gérer l'incertitude et adresse ainsi une des principales limites du Web sémantique, à savoir l'incapacité de représenter et raisonner sur des informations incertaines. Les résultats de cette thèse ont permis de mettre en évidence les avantages des technologies sémantiques par rapport aux approches basées sur de l'apprentissage machine (machine learning) ou fouille de données (data mining). Les travaux en lien avec cette thèse s'inscrivent dans l'axe "Sensibilité au contexte de l'utilisateur" de mes orientations de recherche. J'ai pour ce faire employé des compétences acquises durant ma thèse de doctorat (notamment les modélisations du contexte, calcul de pertinence entre une requête utilisateur et un service Web).
Titre de la thèse (EN)	Enhancing Ontology-Based User Profiling by Uncertainty Handling - An Application to the Digital Advertising Domain	
Date Début – Date Fin	01/10/2012 – 25/03/2016	
% d'encadrement	50%	
Nom et % d'encadrement du directeur	NICOLLE Christophe 50%	
Type de financement	CIFRE (entreprise ezakus, aujourd'hui rachetée par NP6)	
Devenir du doctorant	A signé un contrat de travail à durée indéterminée avec la bibliothèque technique allemande (TIB). Nous continuons d'échanger, et nous souhaitons monter un consortium commun afin de répondre à un appel H2020 ciblant soit le profilage utilisateurs, soit la sensibilité au contexte de l'utilisateur.	
NOM Prénom	**MENDES DE FARIAS Tarcisio**	A travers ce co-encadrement doctoral j'ai pu avancer mes travaux sur l'interopérabilité entre Système d'Information par le biais d'ontologies. Ces travaux sont appliqués ici dans le domaine du bâtiment intelligent, plus particulièrement dans le domaine du BIM (Building Information Modeling). En intégrant les modèles sous-jacents de différents standards, nous avons implémenté une architecture fédérée, faiblement couplée, permettant l'interopérabilité entre ces modèles. Cette architecture n'a pas été spécifiée uniquement dans un contexte BIM ; elle a été définie avec un haut niveau d'abstraction, pouvant être appliquée à tout cas d'utilisation où est défini un alignement entre ontologies à base de règles logiques et
Titre de la thèse (EN)	FOWLA, a Federated Architecture for Ontology Interoperability - An Application to the Building Information Modeling	
Date Début – Date Fin	01/10/2013 – 31/10/2016	
% d'encadrement	50%	

[4] http://www.pr-owl.org

Nom et % d'encadrement du directeur	NICOLLE Christophe 50%	lorsque des requêtes doivent être exécutées au-dessus. Notre approche de fédération d'ontologies permet de réduire de manière drastique le temps d'exécution des requêtes (plus de 80%), sans "perdre" de résultats. Notre principale contribution sous-jacente consiste en un module sélecteur de règles logiques, qui, par rapport à une requête SPARQL donnée, permet de sélectionner uniquement les règles pertinentes à exécuter. Les autres règles sont ignorées, le raisonneur étant uniquement appelée pour les inférences déclenchées par ce sous-ensemble de règles. Les travaux en lien avec cette thèse s'inscrivent dans l'axe "Interopérabilité des Systèmes d'Information" de mes orientations de recherche.
Type de financement	CIFRE (entreprise Active3D aujourd'hui filiale de Sopra Steria)	
Devenir du doctorant	Le doctorant a été embauché par Dassault Systems, à Aix-en-Provence, en CDI, dès la fin de sa thèse. Loin du monde de la recherche (implémentation IFC4), le doctorant a décidé de démissionner. Il est actuellement en postdoc à l'Université de Lausanne / SIB Swiss Institute of Bioinformatics.	
NOM Prénom	**PRUDHOMME Claire**	(*Thèse en cours*) Ces travaux de thèse traitent de la conception d'un SIG sémantique, puis de son implémentation dans un contexte de gestion de crise. Ces travaux sont réalisés en partenariat avec l'Institut i3mainz à Mayence, en Allemagne. L'idée générale est d'utiliser une approche à base de systèmes multi-agents pour pouvoir évaluer des plans de réponse à exécuter lors d'inondations. Pour ce faire, la ville de Cologne (Allemagne) est partenaire du projet. Les recherches effectuées dans ce contexte ont donné naissance à plusieurs publications, chacune d'entre elles présentant une contribution de la thèse : annotation sémantique de données géographiques, intégration de ces données géographiques réparties, évaluation de la notion de "qualité de données géographiques". Actuellement nous travaillons à la spécification formelle de l'approche globale, intégrant les différentes contributions présentées jusqu'ici. Les travaux menés dans le cadre de cette thèse s'inscrivent dans mon axe de recherche traitant du web sémantique spatial.
Titre de la thèse (EN)	Semantic GIS for Disaster Management	
Date Début – Date Fin	01/10/2015 – (en cours)	
% d'encadrement	30%	
Nom et % d'encadrement du directeur	CRUZ Christophe 40% BOOCHS Frank 30%	
Type de financement	Salaire de la doctorante financé par l'Institut i3Mainz de Mayence (Allemagne).	
NOM Prénom	**ARSLAN Muhammad**	(*Thèse en cours*) Cette thèse s'inscrit dans la dernière réorientation de mes axes de recherches, afin d'adresser le domaine des capteurs. L'idée est de permettre la contextualisation des flux de données obtenues à partir de capteurs, en utilisant les technologies sémantiques et ce à différentes échelles : niveau bâtiment, niveau urbain et niveau territoire géographique. Dans un premier temps, nous adressons l'exploitation de données capteurs remontées dans un
Titre de la thèse (EN)	Semantic architecture for Intelligent, Context-aware Sensors in the Built Environment	
Date Début – Date Fin	13/02/2017 – (en cours)	
% d'encadrement	30%	

Nom et % d'encadrement du directeur	CRUZ Christophe 40% GINHAC Dominique 30%	contexte chantier. Ceci est à mettre en lien avec les problématiques du projet ANR McBIM. Par la suite, l'idée est de permettre à des gestionnaires de patrimoine d'avoir une vision d'ensemble allant au-delà du simple bâtiment. Les verrous recherche concernent l'annotation sémantique de sources de données "à la volée" (c'est-à-dire sans scénario défini au préalable). A terme, je souhaiterais orienter les recherches vers l'étude d'événements "cascade" ou événements déclenchés par d'autres événements.
Type de financement	ED SPIM	

2.3.1.2 Encadrement scientifique

2.3.1.2.1 Vue synthétique de l'encadrement scientifique

Autres encadrements	Année universitaire	Nombre	Etudiant(e)s et sujet(s)
Probatoires CNAM	2013-2014	2	Eric GUILLON – Big Data et entreprises Fabien CHAINTREUIL – NoSQL
Thèses de Master	2018-2019	1	Magali BARRUE BASSIN – Réflexion sur une ontologie commune au SIG et au BIM pour les infrastructures : les prémices de la convergence (entreprise eGIS)
Masters 2 Recherche	2015-2016	1	Youssef KERYAKOS – Techniques de positionnement indoor pour la navigation pédestre
	2014-2015	2	Ali ALMOUKAHAL - Smart Cities and Big Data – Context, Challenges and Applications Jiras BIOKOU - Indoor Positioning Techniques: Types, Precision, Use / Application
Projets de fin d'études	2017-2018	3	Antonin ANNET – Développement Web (entreprise Trouve Ton Transport) Jimmy GROSSIORD / Jonathan GERARD – Amélioration application ESIapp
	2016-2017	1	Antoine CANDY – Esiapp (développement d'une application cross-plateformes)
Stages niveau Licence	Au moins 2 par an, et ce depuis 2013 (environ 12 étudiants encadrés).		
Stages niveau Master 1	2017 – 2018	2	Vincent IMBEAUX / Ali Osman SOYKURT – Développement ESIapp
	2012 – 2013	1	Jordan SIMONOT – Adaptation du standard IFC en ontologie OWL
Stages niveau Master 2	2018-2019	2	Benoît BARNOUX – Mise en place d'une solution SALM (Security Audit Logging and Monitoring) (entreprise Mercedes Benz) Ali Osman SOYKURT – Conception et réalisation d'une solution serveur auquel seraient rattachés des mobiles sous Android via USB (entreprise Viveris)
	Depuis 2013	12	3 en 2017-2018, 2 en 2016-2017, 2 en 2015-2016, 3 en 2014-2015 et 2 en 2013-2014

2.3.1.2.2 Vue analytique de l'encadrement scientifique

Le tableau ci-dessous fournit des détails complémentaires concernant les différents encadrements CNAM, Master Recherche ou encore de thèse professionnelle que j'ai effectués jusqu'ici.

Index	Informations générales	
1	NOM	BARRUE BASSIN
	Prénom	Magali
	Type encadrement	Thèse de Master (Thèse Professionnelle) effectuée dans le cadre du Master Spécialisé BIM de l'ENPS (Ecole Nationale des Ponts et Chaussées) – mémoire de fin d'études, niveau bac+6
	Titre	Réflexion sur une ontologie commune au SIG et au BIM pour les infrastructures - les prémices et prémisses de la convergence
	Date Début	01/03/2018
	Date Fin	31/12/2018
	% encadrement	100%
2	NOM	CHAINTREUIL
	Prénom	Fabien
	Type encadrement	Probatoire CNAM
	Titre	NoSQL
	Date Soutenance	23/06/2014
	% encadrement	50%
3	NOM	GUILLON
	Prénom	Eric
	Type encadrement	Probatoire CNAM
	Titre	Big Data et entreprises
	Date Soutenance	17/06/2013
	% encadrement	100%
4	NOM	BIOKOU
	Prénom	Jiras
	Type encadrement	Master 2 Recherche
	Titre	Indoor Positioning Techniques: Types, Precision, Use / Application
	Date Début	01/10/2014
	Date Fin	07/01/2015
	% encadrement	100%
5	NOM	KERYAKOS
	Prénom	Youssef
	Type encadrement	Master 2 Recherche
	Titre	Indoor Positioning Techniques for Pedestrian Navigation
	Date Début	01/10/2015
	Date Fin	29/01/2016
	% encadrement	100%
6	NOM	ALMOUKAHAL
	Prénom	Ali
	Type encadrement	Master 2 Recherche
	Titre	Smart Cities and Big Data – Context, Challenges and Applications
	Date Début	01/10/2014
	Date Fin	07/01/2015
	% encadrement	100%

2.3.2 Projets de recherche

2.3.2.1 Projets en cours

2.3.2.1.1 Projet ANR McBIM (Matière Communicante au service du BIM)

Date début	**Durée**	**Appel à projets**	**Projet**	**Montant de l'aide**	**Mon rôle dans le projet**
01/01/2018	42 mois	ANR AAPG 2017	McBIM (Matière Communicante au service BIM)	124200€	Responsable scientifique pour l'UBFC Gestion administrative du projet Coordination de l'équipe UBFC Responsable de l'embauche d'un chercheur postdoc Rédaction de livrables projet

Dans le cadre du projet ANR McBIM, je suis le responsable scientifique pour le partenaire UBFC. Les objectifs que nous poursuivons sont multiples et concernent :

a) l'extraction de connaissances (e.g. données annotées sémantiquement pouvant être interprétées par un ordinateur) des différentes structures en béton composant un bâtiment,
b) l'intégration de ces connaissances dans la maquette numérique du bâtiment, et
c) leur exploitation tout au long du cycle de vie du bâtiment.

Le standard international pour représenter des maquettes numériques est le format IFC (*Industry Foundation Classes*). Depuis 2017, une sérialisation en langage OWL (*Web Ontology Language*) a été définie pour le format IFC. L'approche étudiée dans ce projet souhaite tirer avantage des technologies dites du Web sémantique, et plus particulièrement des données liées, afin :

a) d'augmenter l'interopérabilité entre les différents acteurs métiers intervenant sur une maquette numérique,
b) de permettre de fournir une connaissance plus précise du bâtiment aux gestionnaires de patrimoine (e.g. résistance structurelle au cours du temps).

Plus particulièrement, les aspects suivants seront abordés :

> Etat de l'art du domaine – identification de vocabulaires contrôlés existants permettant d'annoter les données du bâtiment, ifcOWL, travaux de l'OGC
> Définition du modèle de connaissance minimal permettant l'interopérabilité entre acteurs
> Selon les "jargons métiers" utilisés ou rencontrés, alignement de concepts spécifiques par rapport la terminologie IFC
> Selon les données remontées par le système, identification des propriétés IFC associées
> Définition de règles de gestion permettant de lancer des alertes automatiquement (e.g. suite à une remontée d'informations il faut être capable d'identifier si quelque chose ne va pas).
> Développement d'une plateforme permettant le suivi et la gestion des messages remontés, ainsi que leur intégration par rapport à une maquette numérique existante

A ce titre, je suis responsable de l'embauche d'un(e) chercheur(e) postdoc, sur un contrat à durée déterminée de 18 mois (rédaction de l'offre d'emploi EN/FR, publication sur des sites d'emploi e.g. euraxess-jobs, abg et campusfrance.fr, organisation d'entretiens avec les candidats). Mes autres collègues intervenant dans ce projet sont Monsieur Dominique GINHAC (PR) et Monsieur Wahabou ABDOU (MCF).

2.3.2.1.2 Projet DATAVIEW

Date début	**Durée**	**Projet**	**Partenaires**	**Mon rôle dans le projet**
01/04/2018	24 mois	DATAVIEW Serveur de raisonnement urbain	NF4/NOBATEK EEGLE	Encadrement d'un(e) chercheur(e) postdoc Conduite de projet Formation industrielle de 2j à dispenser

Dans un contexte BIM, la Stratégie française pour les actions de pré-normalisation et normalisation BIM appliquées au bâtiment (Plan Transition Numérique Bâtiment, Feuille de route Normalisation, Avril 2017)

a identifié les données liées en tant que "seule solution technique à la connexion et l'interaction des données hétérogènes manipulées dans un projet BIM". Suite à la publication de cette feuille de route, l'entreprise NOBATEK nous a contactés, le Professeur NICOLLE et moi-même, afin de monter ce projet de recherche. L'objectif poursuivi concerne l'implémentation de mécanismes de fédération d'ontologies dédiés à la modélisation des informations urbaines.

Ce projet est porté par INEF4/Nobatek en partenariat avec la société EEGLE, la Chaire UPPA et le laboratoire LE2I. INEF4/Nobatek nous a sollicités pour intervenir dans l'encadrement scientifique d'un postdoc sur une durée de 24 mois.

Plus particulièrement, il est prévu au cours du programme de collaboration que je dispense une formation aux partenaires du projet. Cette formation concernera le transfert aux industriels de compétences sur la partie fondamentale du projet et du contexte de recherche associé, ainsi que sur la partie applicative pour une prise en main de la preuve de concept réalisée. Cette formation d'une durée de deux jours sera réalisée dans les locaux de NF4/NOBATEK et des supports de formation seront fournis.

2.3.2.1.3 Projet I-SITE UBFC HERMES

Date début (prévue)	**Durée**	**Appel à projets**	**Projet**	**Montant de l'aide**	**Mon rôle dans le projet**
01/09/2018	36 mois	Projet de recherche "Blanc" soumis dans le cadre de l'appel à projets i-Site UBFC 2017	HERMES (spatiotemporal semantics and logical knowledge description of mecHanical objEcts in the era of 4D pRinting and programmable Matter for nExt-generation of CAD systemS	150000€	Co-encadrement d'un doctorant à recruter Participation aux travaux du WP2 "Ontology development and reasoning layers"

Ce projet traite des problématiques théoriques dédiées à la conception pour la fabrication additive 4D, et vise une approche à base de technologies sémantiques. Ma participation à ce projet s'inscrit dans la nouvelle orientation de mes recherches adressant les capteurs et les processus industriels. Ce projet a été sélectionné en mars 2018, dans le cadre du deuxième appel à projets I-SITE porté par l'UBFC. Ce projet regroupe des personnels de l'UTBM (Frédéric DEMOLY, Samuel GOMES) et de l'UB (Christophe CRUZ et moi-même).

2.3.2.2 Réponses à des appels à projets

J'ai rédigé plusieurs propositions de projets de recherche nationaux et internationaux, et j'ai activement participé à leur soumission :

Année	**Appel à projets**	**Projet soumis**	**Partenaires**	**Rôle dans le projet**
2018	ANR AAP Générique 2018	RIAAC (Réseaux Intelligents et Autnomes pour l'Auto-consommation Collective)	Entreprises : CSTB, Sunchain, CATIE, INEF4/Nobatek	Responsable scientifique pour l'UBFC Rédaction de la contribution du LE2I
2017	ANR AAP Générique 2017	McBIM (Matière Communicante au service du BIM)	Académiques : Université de Lorraine (CRAN Nancy), CNRS Midi-Pyrénées (LAAS Toulouse) Entreprise : FINAO (360SmartConnect)	Responsable scientifique pour l'UBFC Rédaction de la contribution du LE2I Choix de l'équipe LE2I

2017	H2020 ICT-16-2017 Big data PPP - research addressing main technology challenges of the data economy	BigSTEP Big Semantic spatio-Temporal data for built Environment Processes ID : 780825	Académiques : Universiteit Gent (Belgique), Technische Universität Darmstadt TUD et Hochschule Mainz (Allemagne), Teesside University (Royaume-Uni), National University of Ireland, Galway (Irelande) Entreprises : Telefónica Germany Next GmbH, IBM Israel, Israel Electric Corporation Limited et Accenture Technology Labs	Coordinateur du projet Montage du consortium (9 partenaires internationaux) Montage et soumission du dossier (70p) Organisation des réunions de consortium (à distance et en présentiel) Création du logo du projet Montage du budget (5M€) Définition de la planification (tâches, work packages, delivrables, milestones, risques, etc.)
2016	ANR MRSEI (Montage de Réseaux Scientifiques Européens et Internationaux)	Données massives sémantiques et spatio-temporelles pour les processus liés à l'environnement bâti	Allemagne : Metasonic GmbH et i3M (Institut für Raumbezogene Informations- und Messtechnik) Belgique : Ghent University Espagne : OEG (Ontology Engineering Group) Irelande : Insight Centre for Data Analytics Pays-Bas : TUE (Eindhoven University of Technology) Portugal : ISEP (Instituto Superior de Engenharia do Porto)	Montage du dossier Montage du consortium
2015	ERASMUS+ Knowledge Alliance Key Action 2: Cooperation for Innovation and the Exchange of Good Practices	PYPICS (Photofits for Young People in Information and Computer Science Careers)	Université de Porto (Portugal), Université de Bourgogne (France), Académie d'Etudes Economiques de Bucarest (Roumanie), Technical University of Denmark, Madrid Polytechnical University (Espagne), entreprise Portugal Telecom (Portugal), entreprise SONAE (Portugal), association DevAcademy (Roumanie)	Responsable scientifique UB Montage du consortium Rédaction de plusieurs parties du dossier
2014	ANR Défi 7 Société de l'information et de la communication	NeuroConsoTex (Neurocognitive profiling of purchase behaviours in the field of fashion industry)	IFTH (Institut Français du Textile et de l'Habillement), IFM (Institut Français de la Mode), IME (Institut de Médecine Environnementale), équipe OUN, équipe CHECKSEM	Montage du consortium Rédaction de la réponse à appel à projets Responsable de la soumission du projet
2014	BQR / Appel à projets Bourgogne Franche-Comté	COCRICUS (COmmunication de CRIse en Contexte Urbain Sémantique)	Equipe OUN, laboratoire ELLIADD, Université de Franche-Comté	Montage du consortium Rédaction de la réponse à appel à projets Responsable de la soumission du projet
2013	ANR France – Roumanie Défi 7 Société de l'information	SEMACRIS (Semantic-based communication system for a	1) Académie d'Etudes Economiques de Bucarest, Roumanie ;	Montage du consortium Rédaction de la réponse à appel à projets

	et de la communication / 7.2 Sciences et technologies numériques	post-crisis disaster situation)	2) Equipe OUN, laboratoire ELLIADD, Université de Franche-Comté, Besançon, France ; 3) Equipe Checksem, laboratoire LE2I, Université de Bourgogne, Dijon, France.	Responsable de la soumission du projet
2013	ULYSSES Partenariat Hubert Curien (PHC) franco-irlandais Domaine : Sciences et technologies de l'information et de la communication	Technologies du data mining et du Web sémantique à l'aide des expertises médico-légales numériques et de la cybercriminalité	PCRG (Parallel Computational Research Group) de l'UCD (University College Dublin)	Rédaction de la réponse à appel à projets Diffusion de l'offre de poste Entretiens avec les candidats

2.4 Evaluation de la recherche

2.4.1 Participation à des comités de programmes de conférences nationales et internationales

Membre de comités de programmes de conférences internationales		
Conférence	**Rôle(s)**	**Contributions**
IEEE SITIS (International Conference on Signal-Image Technology & Internet-Based Systems)	> Responsable du track LHNA (Large, Heterogeneous Networks and their Applications) à partir de 2016 > Co-Responsable du Workshop "Big Data meets Cloud and Virtualized Environments" (BigCVEn) – depuis 2014 et jusqu'en 2017 > Membre comité de programme de la conférence, à partir de 2012 > Membre comité de programme pour le track I-WeCA (Intelligent Web Computing and Applications) en 2013 > Membre comité de programme pour le track "Internet Based Computing and Systems" en 2012 > Membre comité de programme du Workshop "Web based and Distributed Information Systems" en 2012	2017 – 15 articles révisés (1 pour BigCVEn, 14 pour LHNA), chair pour 2 sessions 2016 – 19 articles révisés (16 pour LHNA, 3 pour BigCVEn), chair pour 2 sessions 2015 – 8 articles révisés (4 pour BigCVEn et 4 pour I-WeCA), chair pour 4 sessions 2014 – 8 articles révisés 2013 – 4 articles révisés 2012 – 8 articles révisés
ISWC (International Semantic Web Conference)	> Membre comité de programme Research Track depuis 2017 > En 2017, j'ai été nommée publiquement lors de la cérémonie de fermeture et reçu un certificat me remerciant pour mon investissement au-delà de la normale	2018 – 2 articles révisés 2017 – 6 articles révisés
ESWC (Extended Semantic Web Conference)	> Membre comité de programme "ESWC Research Track" depuis 2017 > Membre du comité de programme de la conférence depuis 2016	2018 – 4 articles révisés 2017 – 6 articles révisés
LDAC (Linked Data for Architecture and Construction)	> Membre du comité de programme de la conférence depuis 2015	2018 – 1 article révisé 2017 – 2 articles révisés 2016 – 2 articles révisés 2015 – 1 article révisé

KMIS (International Conference on Knowledge Management and Information Sharing)	> Membre du comité de programme de la conférence depuis 2016	2018 – 2 articles révisés 2017 – 2 articles révisés 2016 – 2 articles révisés
IEEE IoP (Internet of People)	> Responsable de la publicité (*Publicity chair*) en 2016	Pas d'articles révisés
MCCT International Conference on Modern Communication & Computing Technologies	> Membre comité de programme en 2016 et 2014	2016 – 2 articles révisés 2014 – 2 articles révisés
On The Move federated conferences and workshops	> Membre du comité de programme du workshop Meta4es depuis 2013	2015 – 1 article révisé 2013 – 2 articles révisés
Comités de programme de conférences nationales		
Conférence	**Rôle**	**Contributions**
JFO (Journées Francophones sur les Ontologies)	> Membre du comité de programme à partir de 2016	2016 – 1 article révisé

2.4.2 Comités éditoriaux de revues internationales

Revue	Editeur	Impact factor / Quartile	Rôle
Automation in Construction	Elsevier	2.919 / Q1 en 2017	2018 – 2 articles révisés 2017 – 2 articles révisés
Advanced Engineering Informatics	Elsevier	2.680 / Q1 en 2017	2018 – 1 article révisé
Journal of Data and Knowledge Engineering	Elsevier	1.694 / Q1 en 2015	2015 – 1 article révisé
Journal of World Wide Web	Springer	1.405 / Q2 en 2014	2014 – 1 article révisé
SoftwareX	Elsevier	SCImago Journal Rank (SJR): 3.724 / Q1 en 2017	2018 – 1 article révisé

2.4.3 Commissions d'évaluation et expertise

> **Avril 2018** – J'ai été sollicitée par la Communauté Européenne en tant qu'expert pour l'évaluation de réponses à appels à projets H2020
  > Evaluation des projets soumis lors de l'appel LC-EEB-02-2018 "Building information modelling adapted to efficient renovation"
  > Responsable de la rédaction du rapport d'évaluation final pour 4 réponses à projets
> Depuis **septembre 2017** – Expert français en normalisation BIM (contrat signé avec l'AFNOR, membre délégué au CEN et à l'ISO depuis)
> **Août 2017** – J'ai été sollicitée par le CSTB pour un atelier de travail de 3 jours, dans leurs locaux à Sophia Antipolis

- Le but de cet atelier de travail est de réfléchir aux différents verrous liés à la composition de règles logiques, tout en préservant leur maintenance et lisibilité, ainsi que les performances du moteur sémantique sous-jacent.
- La prestation d'expertise commandée concernait ma participation aux réunions fixées du 30/08/17 au 01/09/17, avec les objectifs suivants :
  - Présentation des travaux réalisés par Madame Ana Roxin en lien avec le thème du workshop
  - Analyse de l'arrêté du 31 janvier 1986, puis produire quelques exemples de règles extraites de cette analyse (jeu de règles)
  - Validation de l'approche par le biais d'un démonstrateur (maquette numérique test + moteur sémantique + jeu de règles créé)
  - Définition d'une spécification pour la rédaction des règles formelles (e.g. syntaxe à utiliser)
  - Définition des recommandations permettant de traiter les verrous.

- Depuis **juin 2016** – Membre du comité éditorial "Feuille de route normalisation" du PTNB
  - Rédaction d'un rapport de 50 pages présentant un état de l'art des initiatives utilisant les données liées dans le domaine de la construction.
  - Participation aux réunions mensuelles du comité éditorial
  - Représentant de la France aux sommets internationaux buildingSmart International (en tant que membre mediaConstruct)

2.4.4 Rapports d'expertise en vue de transferts technologiques

Rédaction de rapports d'expertise (avec la SATT Grand-Est et notamment la filiale de l'Université de Bourgogne dédiée au transfert industriel, Welience)

Année	Nom entreprise	Rôle(s)
2012	IDStart	Définition du périmètre d'innovation technologique de la plateforme IDStart, en particulier dans le domaine de la gestion sémantique des projets proposés sur la plateforme.
2012	DigitAPP	Réalisation d'une étude sur la modélisation sémantique d'un ERP (Enterprise Ressource Planner)
2013	Webdrone	Analyse des verrous de recherche liés au projet de la société et proposition de solutions d'innovation technologique qui y répondent.

2.4.5 Administration de la science

2.4.5.1 Responsabilité de la logistique organisationnelle de congrès internationaux

Conférence	Rôle(s)
LDAC (Linked Data for Architecture and Construction)	> Memebre comité organisation conférence LDAC (annonce lors de l'édition 2016 que l'édition 2017 se tiendra à Dijon) > Membre comité éditorial pour les actes de la conférence LDAC2018 > Organisation de la conférence LDAC2017 à Dijon > Réalisation site Web http://linkedbuildingdata.net/ldac2017/ > Responsable programme de la conférence > Responsable des conférences invitées (*Keynote*) – Monsieur Christophe CASTAING (eGIS) et Monsieur Benoît VERVANDIER (Groupe Archimen) > Logistique inscription (création et impression de badges), repas (devis traiteur, bons de commande, livraison et ramassage), animations (sorties restaurants prévues chaque soir), hébergement, goodies (en lien avec la boutique de l'uB et constitution de sachets cadeau offerts à chaque participant contenant un bloc-notes, un stylo et une clé USB uB)

IEEE SITIS (International Conference on Signal-Image Technology & Internet-Based Systems), track chair pour LHNA (Large, Heterogenous Networks and their Applications)	> Participation à l'organisation des éditions : > 2015 – chair pour 4 sessions, gestion du comité de programme pour le workshop BigCVEn, révision d'articles, publicité > 2016 – chair pour 2 sessions, gestion du comité de programme pour le track LHNA, révision d'articles, publicité > 2017 – chair pour 2 sessions, gestion du comité de programme pour le track LHNA, révision d'articles, publicité, liaison avec organisateurs locaux, chair de 2 sessions (délivrance de certificats de participation) > 2018 – responsable du track LHNA, gestion du *program commitee*, gestion du processus de revue des articles.

2.4.5.2 Responsabilité de la logistique organisationnelle de manifestations scientifiques nationales

Date(s)	Conférence	Rôle(s)
10 mars 2011	Journée "Géopositionnement et mobilités intelligentes », organisée par le laboratoire SeT (Systèmes et Transports) de l'UTBM	> Membre du comité d'organisation > Réalisation de dépliants > Participation à la définition du programme de la journée > Responsable conférenciers invités
Du 3 au 20 juin 2013	Exposition BeConnected	> j'ai activement participé au montage de l'exposition Objets Connectés à la Maison Régionale de l'Innovation (MRI) à Dijon. > J'ai participé à plusieurs réunions avec des personnels de la MRI > J'ai aussi construit et fourni un dossier complet sur les objets connectés (spécification du contenu de plusieurs panneaux de l'exposition, création de contenus multimédia pour l'exposition, etc.)
26-27 mai 2016	BIMoc (BIM et objets connectés)	> J'ai participé au montage du salon BIMoc, qui a eu lieu le 26 et le 27 mai 2016, à Dijon

2.4.5.3 Responsabilités au sein de son établissement

> Lors des élections UBFC du 22 et 23 mars 2016, j'ai été candidat sur la liste "UBFC : Osons notre avenir" dirigée par Madame Annie VINTER – pour le conseil académique (CAC) et pour le collège B "Autres enseignants-chercheurs, enseignants et personnels assimilés, ou équivalent"

2.4.5.4 Membre actif d'organisations de standardisation nationales et internationales

AFNOR (Association française de normalisation)	
Membre depuis	Septembre 2017
Rôle	> Membre de la commission AFNOR PPBIM "Maquettes numériques dans la construction", plus particulièrement j'interviens dans les groupes d'études GE1 "Feuille de route" et GE3 "Information Delivery Manual", avec les missions suivantes: > Cartographie des normes du domaine et champs d'application, élaboré à partir de la feuille de route PTNB (à la rédaction de laquelle j'ai participé en tant que membre du comité éditorial) > Identification des points de vigilances dans chacun des projets de normalisation > Participation à des réunions mensuelles, en présentiel ou à distance > Révision de proposition de standards européens et internationaux, formulation de commentaires CEN/ISO > Participation à l'élaboration d'une position française quant à l'ordonnancement des travaux CEN et ISO

	> Participation aux travaux du CEN/TC 442 WG 5 (TG Stratégie et TG rôle horizontal) > Participation aux travaux de l'ISO/TC 59 SC13 préparation TF 02 – stratégie
CEN (Comité Européen de Normalisation)	
Membre délégué depuis	Septembre 2017
Rôle	Intervient dans les discussions du comité technique CEN/TC442, notamment dans les groupes de travail WG2 "Formats d'échange d'information", WG3 "Processus" et WG4 "Dictionnaires de propriétés" Participe en tant qu'expert français dans les discussions sur les propositions suivantes: > PWI 442007 (prEN ISO 23386) Building information modelling and other digital processes in Construction - Methodology to describe, author and maintain properties in interconnected dictionaries [issu de la norme française XP 07-150] > PWI 442010 (ISO/NP 23387) Product data templates, for products and systems used in construction works, stored in a data dictionary framework – Part 1: General concepts, relations, and general structure of product data templates, and how to link the product data templates to Industry Foundation Classes (IFC) > PWI 442008 Product data templates, for products and systems used in construction works, stored in a data dictionary framework – Part 2: Specification of Product data templates based on harmonised technical specifications under the Construction Products Regulation (CPR), and how to link the product data templates to Industry Foundation Classes (IFC)
ISO (International Organization for Standardization)	
Membre depuis	Octobre 2017
Rôle	Membre du comité ISO TC 59/SC 13 "Organization and digitization of information about buildings and civil engineering works, including building information modelling (BIM)" Membre des groupes de travail suivants: > WG 08 "Building information models - Information delivery manual" > WG 13 "Implementation of collaborative working over the asset lifecycle" Membre du comité éditorial pour la norme ISO 21597 "Information container for data drop" - Exchange specification > Réunions hebdomadaires de 2h, sous la direction de Henk Schaap (Convenor Working Group IDM ISO TC 59/SC 13) > Part 1: Container > Part 2: Dynamic semantics Membre du groupe de travail joint entre l'ISO/TC 59/SC 13 et l'ISO/TC 211 sur l'interopérabilité BIM et SIG > ISO TR 23262 "GIS (Geospatial) / BIM interoperability" Intervenant comme expert dans les discussions concernant les propositions de normes suivantes: > ISO/DIS 19650 "Organization of information about construction works — Information management using Building Information Modelling (BIM)" > Part 5: Specification for security-minded building information modelling, digital built environments and smart asset management > NWI (New Work Item) ISO TC59/SC13 N178 "Framework for Provision of Guidance on Building Information Modelling"
W3C (World Wide Web Consortium)	

Membre depuis	9 Decembre 2014
Rôle	Membre du "Linked Building Data" Community Group > Responsable du sous-groupe "Ontology Alignment Group (OAG)" en charge de la définition d'alignements entre vocabulaires contrôlés standard existants dans le domaine AEC (Architecture, Engineering, Construction) > Je participe au montage du futur "Buildings on the Web" Working Group qui sera présenté au prochain TPAC du W3C, le lundi 22 octobre 2018 à Lyon, de 8h30 à 10h30.
buildingSmart International	
Membre depuis	Septembre 2012
Rôle	> Membre de la Technical Room – maintenance des principaux standards IFC (IDM, IFC, MVD), amélioration de la spécification IFC, contribution à la définition de IFC Alignments > Réunions mensuelles à distance > Membre du Linked Data Working Group (LDWG) – contribution à l'amélioration d'ifcOWL, définition de liens vers le buildingSmart Data Dictionnary (bSDD) > Responsable des interactions avec la Product Room > Membre de buildingSmart Forums – communauté d'utilisateurs de buildingSmart visant l'amélioration de l'utilisation, du développement et de l'implémentation des standards internationaux définis par buildingSmart, ainsi que des technologies sous-jacentes > Mes contributions visent l'application des technologies sémantiques, plus particulièrement des données liées, afin de notamment faciliter le traitement de maquettes numériques basées sur ifcOWL. Je m'intéresse aussi à la définition de MVD (Model View Definition) par le biais de règles logiques manipulant des concepts présents dans ifcOWL ou d'autres vocabulaires.
OGC (Open Geospatial Consortium)	
Membre depuis	Decembre 2012
Rôle	Membre du Geosemantics DWG Mise en relation du CSTB (Centre Scientifique et Technique du Bâtiment) avec l'OGC

2.4.5.5 Membre de groupements de recherche nationaux

GdR MaDICS 3708 (Masses de Données, Informations et Connaissances en Sciences)	
Membre depuis	Juillet 2015
Rôle	Participation à la rédaction de la proposition d'atelier pour l'Assemblée Générale constitutive du GdR MaDICS de juin 2015 > Titre "Vers une compréhension des données hétérogènes par approches sémantiques" > Résumé: "L'objectif général de l'atelier est d'étudier comment les approches sémantiques permettent d'améliorer la compréhension des masses de données hétérogènes, fortement impactées par les deux "V" que sont la Véracité et la Valeur. L'atelier proposé vise à fédérer dans une approche interdisciplinaire les chercheurs exploitant les données complexes, dont les modèles de représentation de la réalité se fondent sur la modélisation sémantique; et pour lesquels les chercheurs en informatique peuvent proposer des modèles, adapter des algorithmes et des outils de raisonnement."
GdR MAGIS 2340 (Méthodes et Applications pour la Géomatique et l'Information Spatiale)	
Membre depuis	Septembre 2013

2.5 Diffusion des travaux (rayonnement et vulgarisation)

Au moment de la rédaction de ce document, mes publications totalisent 51 éléments, parmi lesquels : 9 articles dans des revues internationales avec comités de lecture (18%), 28 articles dans des conférences internationales avec comités de lecture et actes (54%), 2 articles dans des conférences nationale (4%) et 7 chapitres d'ouvrages (14%).

Le synopsis ci-dessous se base sur des données de mon profil Google Scholar et illustre ma volonté d'avoir un rythme de publication constant et soutenu.

	ACTUELLEMENT	EN JUIN 2018	EN 2016	EN 2013	EN 2011
CITATIONS	351	314	165	67	20
H-INDEX	8	7	7	4	4
I10-INDEX	7	6	6	1	1

2.5.1 Publications majeures

Le tableau ci-dessous présente une sélection de mes publications majeures. Je les ai choisies pour deux raisons : d'une part, elles témoignent de ma volonté de viser les revues internationales (si possible indexées ISI Web of Science) et les conférences internationales sélectives ; d'autre part, elles sont représentatives de mes collaborations avec des chercheurs internationaux. Pour chacune des publications sélectionnées, je présente son contexte d'écriture, ma contribution ainsi que leur résumé.

[RIS5] Tarcisio Mendes De Farias, *Ana Roxin*, Christophe Nicolle. **SWRL rule-selection methodology for ontology interoperability**. Data & Knowledge Engineering, 105, pp.53-72. IF en 2015: 1.12. Taux d'acceptation pour l'appel considéré: 7 %. Rang journal SCImago en 2015: Q1	
Abstract	Data interoperability represents a great challenge for today's enterprises. Indeed, they use various information systems, each relying on several different models for data representation. Ontologies and notably ontology matching have been recognized as interesting approaches for solving the data interoperability problem. In this paper, we focus on improving the performance of queries addressed over ontology alignments expressed through SWRL rules. Indeed, when considering the context of executing queries over complex and numerous alignments, the number of SWRL rules highly impacts the query execution time. Moreover, when hybrid or backward-chaining reasoning is applied, the query execution time may grow exponentially. Still, the reasoners involved deliver performant results (in terms of execution time) when applied over reduced and simpler rule sets. Based on this statement, and to address the issue of improving the query execution time, we describe a novel approach that allows, for a given query, to ignore unnecessary rules. The proposed Rule Selector (RS) is a middleware between the considered systems and the reasoner present on the triple store side. Through the benchmarks realized we prove that our approach allows considerably minimizing query execution time.
Placement dans le contexte recherche	Ce travail adresse la problématique de l'interopérabilité des Systèmes d'Information à travers des alignements d'ontologies exprimés sous la forme de règles logiques SWRL (Semantic Web Rule Language). Plus particulièrement, nous nous intéressons ici à l'amélioration du temps de réponse des requêtes SPARQL (Simple Protocol and Query Language) exécutées sur ce type d'alignement d'ontologies. Initialement conçue pour le domaine du la modélisation d'informations du bâtiment, notre approche a été adaptée afin de pouvoir être appliquée dans tout contexte métier nécessitant une interopérabilité entre modèles. Le principal avantage de notre approche est d'être flexible et de permettre de répondre à des requêtes plus rapidement que d'autres approches similaires (ceci est prouvé à travers des résultats numériques).
[CI14] Anett Hoppe, *Ana Roxin*, Christophe Nicolle. **Automatic User Profile Mapping To Marketing Segments In A Big Data Context**. 14th International Conference on Informatics in Economy (IE 2015), Apr 2015, Bucharest, Romania. Conférence IE2015 membre de Web of Science Core Collection	
Abstract	Within the discussion about the analysis methods for Big Data contexts, semantic technologies often get discarded for reasons of efficiency. While machine learning and statistics are

	known to have shortcomings when handling natural language, their advantages in terms of performance outweigh potential concerns. We argue that even when handling vast amounts of data, the usage of semantic technologies can be profitable and demonstrate this by developing an ontology-based system for automatically mapping user profiles to pre-defined marketing segments.
Placement dans le contexte recherche	Les travaux présentés dans cet article représentent la synthèse des années de recherche précédentes. Nous y présentons notre approche permettant de profiler un utilisateur selon des segments marketing, en utilisant principalement des données de navigation (e.g. cookies). Nos travaux sont appliqués dans le domaine du marketing en ligne, avec la définition d'interfaces intuitives développées spécialement pour ce contexte d'application.
[RIS3] Pieter Pauwels, Tarcisio Mendes De Farias, Chi Zhang, *Ana Roxin*, Jakob Beetz, et al. **A performance benchmark over semantic rule checking approaches in construction industry**. Advanced Engineering Informatics, Elsevier, 2017, 33, pp.68-88. DOI: 10.1016/j.aei.2017.05.001 Rang journal SCImago: Q1	
Abstract	As more and more architectural design and construction data is represented in the Resource Description Framework (RDF) data model, it makes sense to take advantage of the logical basis of RDF and realise a semantic rule-checking process as it is currently not available in architectural design and construction industry. The argument for such a semantic rule-checking process has been made a number of times by now. However, there are a number of strategies and approaches that can be followed regarding the realisation of such a rule-checking process, even when limiting to the use of semantic web technologies. In this article, we will outline three reference rule-checking approaches that have been reported earlier for semantic rule-checking in the AEC domain. Each of these approaches has its advantages and disadvantages. A criterion that is tremendously important to allow adoption and uptake of such semantic rule checking approaches, is performance. Hence, this article provides an overview of our collaborative test results in order to obtain a performance benchmark for these approaches. The benchmark is discussed in this paper, in addition to a brief documentation of the actual rule-checking approaches. Furthermore, we give an indication of the main features and decisions that impact performance for each of these approaches, so that system developers in construction industry can make an informed choice when deciding for one of the documented rule-checking approaches, perhaps even sacrificing part of the performance.
Placement dans le contexte recherche	Etant donné l'intérêt grandissant durant les dernières années pour les approches de vérification de maquettes numériques reposant sur des technologies du Web sémantique, cet article représente la première initiative d'une comparaison de ces approches. Au moment de la rédaction de cet article, il n'existait aucune référence concernant le niveau d'efficacité qu'une telle approche doit atteindre lorsqu'implémentée. Cet article présente donc trois manières d'implémenter ce type d'approche et fournit des éléments de comparaison (quantitative et qualitative) par rapport à ces approches. Chacune des implémentations présentées est spécifique à une équipe de recherche internationale. L'idée de cet article est de présenter les approches choisies par chacune des équipes, avec les raisons ayant justifié ce choix. Les tableaux comparant les résultats atteints avec chacune des approches doit aider les futurs chercheurs dans le choix de leur implémentation.
Tarcisio Mendes De Farias, *Ana Roxin*, Christophe Nicolle. **A rule-based methodology to extract building model views**. Automation in Construction, Elsevier, 2018, pp.214-229. Rang journal SCImago: Q1	
Abstract	In this paper, we present a novel approach called IfcView that relies on Semantic Web technologies for creating building views. To do so, we consider an ifcOWL ontology proposed by buildingSMART. The ifcOWL is an Industry Foundation Classes (IFC) based ontology. By combining the ifcOWL ontology with logical rules (expressed in Semantic Web Rule Language, SWRL), we demonstrate through several case studies that our approach can perform a more intuitive and flexible extraction of building views when compared to the Model View Definition (MVD) approach. This is because our rule-based approach dynamically creates sub-graphs (i.e. views) by specifying the IFC elements to extract as Globally Unique Identifiers (GUID), relationships or entities. Another benefit of our approach is the fact that it simplifies the maintenance and definition of building views. Once our rule-

	based system extracts such a building view (i.e. sub-graph), this view can be exported by using STEP (STandard for the Exchange of Product) or Turtle (a Resource Description Framework (RDF) syntax) formats.
Placement dans le contexte recherche	Cet article présente notre approche d'extraction de vues IFC en utilisant les technologies sémantiques. Cette approche se veut une alternative à l'approche traditionnelle promue par buildingSmart International (bSI), à savoir l'approche MVD (Model View Definition). A partir d'un fichier IFC STEP, nous avons développé un algorithme en Java permettant de "traduire" ce fichier au format OWL, en respectant l'ontologie ifcOWL, telle que soutenue par bSI. Ce fichier OWL est stocké dans un magasin de triplets en vue d'être requêté. Toutefois, le langage de requêtes pour le Web sémantique (e.g. SPARQL) nécessite un apprentissage conséquent. Dès lors, nous avons développé un deuxième algorithme qui, sur la base d'éléments fournis par l'utilisateur, compose automatiquement la requête SPARQL associée. L'utilisateur final n'a ainsi pas besoin d'apprendre un nouveau langage de requêtes : il lui suffit de préciser le nom des concepts et/ou des relations et/ou l'identifiant des éléments à extraire (e.g. GUID). Une fois la requête SPARQL générée et exécutée, les résultats sont passés en entrée à un troisième algorithme qui est reponsable de générer le fichier IFC STEP contenant seulement les éléments retournés par la requête. A travers trois cas d'étude, nous prouvons que cette approche a plusieurs avantages lorsque comparée à l'approche MVD (e.g. flexibilité, simplicité). Cet article discute les performances de l'approche ainsi que des améliorations possibles.

2.5.2 Séminaires et conférences "invité"

- 26/03/2018 – Présentation **"Linked (meta)Data – Principles and Applications for BIM"**. Session plénière au buildingSMART International Standards Summit, Espace Grande Arche à La Défense, 26-29 mars 2018, Paris, France.
  - Description de l'événement: "Dans le cadre du sommet de l'openBIM de buildingSMART International, qui rassemblera plus de 300 experts et influenceurs du BIM du monde entier, la plénière publique est ouvert au public le 26 mars 2018: ce sera l'International openBIM day[5]"
  - Présentation disponible en ligne
- 25 – 30/06/2017 – **Invitation et participation au séminaire international Dagsthul 17262** "Federated Semantic Data Management",
  - J'ai rejoint le groupe discutant des cas d'usage et des applications reposant sur une gestion de données sémantiques fédérées. Nous avons cherché à présenter une vision, articuler les avantages et identifier les limites des principales approches dans le domaine du FDSM.
  - Présentation disponible en ligne
  - Le rapport issu de ce séminaire est disponible en ligne.
- 18/05/2017 – Présentation **"Technologies du Web sémantique pour les classifications"** – Invitation au GT "Classifications" de mediaConstruct, dans les locaux de l'association Qualitel, à Paris.
- 04/04/2017 – Présentation jointe avec Matthias Weise (AEC3 Allemagne) : **"Model View Definitions and Linked Data – The MVD Whitepaper"**. Session de la Technical Room, au buildingSMART International Standards Summit, 3-6 avril 2017, Barcelona, Spain.
  - Après une introduction présentant le concept des Model View Definition (MVD) par Matthias Weise, nous avons présenté trois principales approches appliquant les technologies sémantiques au concept MVD (e.g. publication / filtrage / hybride).
  - J'ai présenté une implémentation de l'approche de filtrage par rapport au MVD COBie. Cette approche se base sur mes travaux sur l'ontologie COBieOWL ainsi que sur l'adaptation en tant que règles logiques SWRL des clauses contenues dans le MVD COBie. Sur la base de cette approche, j'ai présenté une généralisation possible, qui permet d'extraire des parties d'un

[5] http://www.mediaconstruct.fr/sinformer/agenda-du-bim/udt_812_param_detail/8940

fichier IFC. Cette généralisation a été l'idée de base pour nos travaux publiés en 2018 sur la définition de vues IFC avec des règles logiques [RIS1].
  - Présentation disponible en ligne

- 29/03/2017 – Présentation "**Les données liées pour le BIM**". Séance "En quoi les normes du BIM sont une aide pour les professionnels ? Les enjeux concrets de la normalisation BIM", bimWorld 2017, 29-30 mars 2017, Espace Grande Arche à La Défense, Paris, France
  - J'ai été invitée par mediaConstruct en tant qu'expert normalisation français pour présenter les réponses apportées par les données liées pour faciliter la manipulation des maquettes numériques de bâtiments
  - Présentation disponible en ligne
- 27/10/2016 – Présentation "**A Linked Data Perspective for BIM**" - séance jointe Technical Room / Building Room / Product Room / Infrastructure Room, au buildingSmart International Standards Summit, 25-29 septembre 2016, Jeju, Corée du Sud.
  - J'ai été invitée pour présenter comment une approche à base de données liées pourrait être appliquée dans le domaine du BIM et quels problèmes elle permettrait d'adresser
  - Présentation disponible en ligne
- 26/10/2016 – Présentation jointe avec Pieter Pauwels (Univ. Ghent, Belgique) – "**Reasoning with rules – Application to N3/EYE and Stardog**". Session de la Regulatory Room, au buildingSMART International Standards Summit, 25-29 septembre 2016, Jeju, Corée du Sud.
  - Nous avons chacun présenté les approches que nous utilisons pour vérifier des maquettes numériques à l'aide de règles logiques. Pieter Pauwels a présenté une approche employant un raisonneur *open source* (e.g. EYE) et une sérialisation des données sémantiques en N3.
  - J'ai présenté une approche basée sur le magasin de triplets Stardog, qui malgré le fait de nécessiter une licence payante, a l'avantage d'être très facile à prendre en main pour des non-spécialistes du Web sémantique.
  - J'ai fait une démonstration en temps-réel de comment une telle approche peut être utilisée.
  - Présentation disponible en ligne
- 26/10/2016 – Présentation "**A Federated Approach for interoperating AEC/FM Ontologies**". Session de la Technical Room, au buildingSMART International Standards Summit, 25-29 septembre 2016, Jeju, Corée du Sud.
  - J'ai présenté notre approche de fédération d'ontologies à base de règles logiques et son application aux deux standards IFC et COBie.
  - Présentation disponible en ligne
- 13/04/2016 – Présentation "**A Semantic Web Approach for Building View Definitions**". Session de la Technical Room session, au buildingSMART International Standards Summit, 11-14 avril 2016, Rotterdam, Pays-Bas.
  - J'ai présenté notre approche permettant de traduire un fichier IFC STEP en ifcOWL, en extraire une sous-partie en spécifiant les concepts/relations à extraire soit par leur nom, soit par leur GUID, puis de générer le fichier IFC STEP résultant.
  - Présentation disponible en ligne
- 12/04/2016 – Présentation "**IfcWoD (Web of Data) – Semantically adapting IFC Model Relations into OWL Properties**". Session de la Technical Room session, au buildingSMART International Standards Summit, 11-14 avril 2016, Rotterdam, Pays-Bas.
  - J'ai présenté notre approche pour appliquer les principes des données liées à l'ontologie ifcOWL, ainsi que les avantages offerts par cette approche par rapport à la modélisation actuelle (requêtes plus simples à composer, plus rapides à exécuter)
  - Présentation disponible en ligne

- 29/02/2012 – Présentation "**Les technologies du Web sémantique pour la mobilité et les réseaux sociaux**". Séminaire invité à l'INRIalpes, pour l'équipe EXMO (actuellement mOeX) du Laboratoire d'informatique de Grenoble (LIG). Invitation de Jérôme Euzenat.
  - Après une brève introduction aux technologies du Web sémantique, j'ai présenté plusieurs approches appliquant ces technologies dans deux domaines : la mobilité et l'interaction sociale. Parmi les outils sémantiques pour la mobilité, j'ai présenté le protocole de découverte pour services Web sémantiques mis au point durant ma thèse de doctorat, ainsi que l'initiative DbPedia Mobile. J'ai aussi présenté les principaux vocabulaires standards permettant d'annoter les interactions entre utilisateurs sur des réseaux sociaux (e.g. foaf, SIOC, SKOS, RDFa, OpenGraph).
  - Présentation disponible en ligne
- 11/02/2011 – Présentation "**Les technologies du Web sémantique pour la modélisation et l'interopérabilité des informations géographiques**". Séminaire invité au Laboratoire TOPO (Geodetic Engineering Laboratory) de l'EPFL. Invitation de Pierre-Yves Gilléron.
  - Après une brève introduction aux technologies du Web sémantique, j'ai présenté plusieurs approches appliquant ces technologies pour la manipulation et l'intégration d'informations géospatiales. J'ai notamment dressé un état de l'art des approches à base de sémantiques pour les informations géographiques (de 2000 à 2010). J'ai aussi présenté comment les technologies du Web sémantique pouvaient aider à la création, la description et l'exécution de géo services (ou services Web manipulant des informations géographiques).
  - Présentation disponible en ligne
- 10/03/2010 – Présentation "**L'information géographique"**. Journée "Mobilités intelligentes et Géopositionnement" organisée à l'UTBM.
  - Après une brève définition de ce qu'on entend par "information géographique", j'ai présenté sous quelles formes ce type d'information pouvait être représenté. J'ai ensuite présenté les différents standards ISO et/ou OGC pour les informations géographiques e.g. SVG (Scalable Vector Graphics), GML (Geography Markup Language) et GDF (Geographic Data Files). J'ai terminé la présentation en discutant des techniques existantes pour collecter des informations géographiques e.g. photogrammétrie, LiDAR.
  - Présentation disponible en ligne
- 04/02/2010 – Présentation "**Protocole de découverte, sensible au contexte, pour les services Web sémantiques**". Journée "Réseaux Grand Est (RGE)" organisée par l'équipe CARTOON du LIFC, à Besançon.
  - Présentation de l'approche conçue durant ma thèse de doctorat, adressant la découverte des services Web sémantiques.

2.5.3 Collaborations internationales

- Participation à l'atelier de travail "**Unfolding Through Time: Digital Innovation in the Built Environment**», organisé par University College London (UCL) plus particulièrement leur Institute for Digital Innovation in the Built Environment. Cet atelier a été financé par l'Ambassade Française à Londres, et s'est tenu du 20 au 21 juillet 2017 à la Bartlett School of Architecture.
  - J'y ai participé avec des d'autres personnes du CSTB (Souheil SOUBRA, Julien SOULA)
  - Le but était d'identifier les prochains appels à projets internationaux et nationaux auxquels nos institutions respectives pouvaient répondre avec des propositions de recherche.
  - Deux axes ont été identifiés, un pour la France (vérification de maquettes numériques) et un pour le Royaume-Uni (les technologies du blockchain en construction).
- A partir de 2014, mise en place d'une collaboration étroite avec l'entreprise Stardog – s'agissant des magasins de triplets très performants, que nous avons testé dans plusieurs approches de recherche, j'ai contacté l'entreprise Stardog afin d'initier une collaboration entre eux et nos chercheurs. Le 30/06/2014 un *memorandum of understanding* a été signé entre l'entreprise et l'Université de Bourgogne, représentée par l'équipe Checksem du laboratoire LE2I. Depuis, je suis

en charge de la relation avec cet industriel. Nous avons notamment remonté plusieurs bugs de l'application, que les développeurs Stardog ont corrigés. Enfin, dans l'article de journal soumis à "Advanced Engineering Informatics" **Erreur ! Source du renvoi introuvable.**, nous présentons des ésultats obtenus par rapport à une implémentation de notre approche sur un magasin de triplets Stardog. Nos résultats sont les meilleurs en termes de temps d'exécution de requêtes et d'exhaustivité des résultats, ceci en partie grâce à la collaboration étroite que j'ai mis en place entre notre équipe de recherche et l'entreprise.

- Collaborations avec l'Académie d'Etudes Economiques de Bucarest (Roumanie), l'Université d'Oradea (Roumanie), l'Université de Technologie de Belfort-Montbéliard (UTBM) et le laboratoire Topométrie de l'Ecole Polytechnique Fédérale de Lausanne (EPFL)
- Signature de plusieurs accords de confidentialité avec des entreprises françaises et étrangères (Danemark, Etats-Unis, Allemagne)

Collaborations avec des chercheurs internationaux :

- Pieter Pauwels – cette collaboration a donné naissance à plusieurs publications jointes [RIS3], [CH4], [CI7]
- Udo Kannengiesser – il s'agit d'une recherche exploratoire sur l'apport des technologies sémantiques dans la modélisation de processus humains dans un contexte BIM (S-BPM contre BPMN) ; cela a donné naissance à une publication jointe [CH5]; au travers des travaux de Madame Claire PRUDHOMME j'ai souhaité réutiliser certains résultats de cette étude.

2.5.4 Collaborations nationales

2.5.4.1 Collaboration avec le CSTB (Centre Scientifique et Technique du Bâtiment)

- Sur demande de Souheil SOUBRA (Directeur Technologies de l'Information du CSBT), j'ai rejoint le comité d'experts normalisation français et participé à la rédaction de la Stratégie française pour les actions de pré-normalisation et normalisation BIM appliquées au bâtiment
  - Signature d'un contrat d'expertise entre la SATT Grand-Est et le CSTB définissant le cadre de ma participation à cette feuille de route
- Par la suite, j'ai rejoint le projet "Numérisation de règles" correspondant à la mise en place d'un démonstrateur d'un service de vérification des exigences technico-réglementaires de maquettes numériques
  - Projet en collaboration avec la Direction Générale de l'Aménagement du Logement et de la Nature (rattachée au Ministère du Logement et de l'Habitat durable)
  - Signature d'un contrat de collaboration avec le CSTB, d'une durée de 18 mois, à partir du 28/08/2017 ; la mission qui m'est affectée dans ce contexte concerne la contribution à l'organisation et la conceptualisation des connaissances ainsi que la formalisation des règles associées aux cas d'usage retenus pour le démonstrateur.
  - Participation à un atelier de travail sur 3 jours, organisé dans les locaux du CSBT à Sophia Antipolis.
  - Cette collaboration a permis la publication deux articles joints : un lors de la conférence annuelle du groupe de travail W3C Linked Building Data [CI1], et un deuxième pour la conférence ECPPM 2018 [CH1].

2.5.4.2 Collaboration avec CERQUAL / Qualitel

- Suite à ma présentation lors du bimWorld en 2017, j'ai été contactée par Monsieur Yannick COTHEREL, Responsable d'activité BIM au sein de la Direction des Etudes et Recherche du groupe CERQUAL.
- Nous avons entamé des discussions sur l'utilisation des technologies sémantiques pour la vérification de maquettes numériques, discussions qui ont donné naissance à trois axes de collaboration :
  - Dans le cadre du GT 3.1 "Les systèmes de classification" de mediaConstruct : participation à la rédaction de l'état de l'art des systèmes de classification existants (soumis pour validation à

buildingSmart International), définition de correspondances sémantiques (semantic mappings) entre certaines classifications.
- Projet de recherche "QualiBIM" – j'ai monté avec l'appui de la SATT Grand-Est, un projet de collaboration scientifique dans le cadre duquel je serais responsable de l'encadrement d'un chercheur postdoc sur une durée de 18 à 24 mois. Dans le cadre de ce projet, CERQUAL souhaite étudier si les technologies sémantiques peuvent aider à la vérification (semi-) automatisée de maquettes numériques renseignées "BIM", et ce dans le cadre du processus de la certification NF Habitat et NH Habitat HQE.
- Projet à venir Boost-Construction : projet plus global visant l'étude de l'impact de l'introduction des technologies sémantiques dans les projets de construction.

2.5.4.3 Collaboration avec Maxime LEFRANCOIS (MINES Saint-Etienne)

- Suite à notre rencontre lors de la formation à l'écriture de projets Horizon 2020[6], nous avons débuté une collaboration étroite notamment :
  - Dans le cadre du projet "Numérisation de règles" avec le CSTB – j'ai invité Monsieur LEFRANCOIS à participer à l'atelier de travail organisé l'année dernière au mois d'août, dans les locaux du CSTB, à Sophia Antipolis. Il a ainsi pu présenter l'approche à base motifs ontologiques (ontology pattern) utilisée dans la conception des ontologies SEAS (Smart-Energy Aware Systems) et nous avons discuté des avantages d'une telle approche pour le projet "Numérisation de règles".
  - Dans le cadre du groupe communautaire W3C "Linked Building Data" où nous intervenons ensemble dans le groupe traitant des alignements entre ontologies (Ontology Alignement Group).

2.5.4.4 Collaboration avec l'équipe OUN (Objets et Usages Numériques) du laboratoire ELLIADD

- Créé en 2012 suite au regroupement des EA 2181 (LASELDI) et EA3187 (ATST-Centre Jacques-Petit), l'EA 4661 ELLIADD (Edition, Littératures, Langages, Informatique, Arts, Didactique, Discours) est une Unité de Recherche de l'Université de Franche-Comté évaluée A par l'AERES pour son projet scientifique.
- A partir de 2013, j'ai mis en place d'une collaboration entre l'équipe Checksem du LE2I et l'équipe OUN du laboratoire ELLIADD.
  - J'ai invité plusieurs membres de l'équipe OUN, à Dijon, pour plusieurs réunions de travail. Le but de ces rencontres était d'identifier les axes de recherche de chaque équipe (Checksem et OUN), afin de dégager des idées de collaboration. Une première réunion a eu lieu le 16/04/2013 sur Dijon.
  - Suite à cette première réunion, un memorandum of understanding (MOU) a été signé entre les 2 équipes. Depuis, nous avons eu plusieurs autres réunions, et nous avons soumis plusieurs projets de recherche et BQR ensemble (listés dans la section "Réponses à des appels à projets" de ce document).

2.5.5 Formation aux industriels

- Courant 2020 – dans le cadre du projet DATAVIEW, je vais être en charge d'une formation à dispenser aux équipes de Nobatek et traitant des technologies sémantiques pour la manipulation et l'intégration d'informations réparties dans le domaine de la construction
  - Formation de 2 jours, prévue dans les locaux de Nobatek sur Bordeaux
- Décembre 2015 – atelier de formation sur 2 jours (14h au total), 8 personnes y ont participé.
  - J'ai été en charge de journées de formation à des industriels bourguignons (Groupe SEB), sur l'ingénierie des connaissances. La formation a eu une durée totale de 14h, réparties sur 2 jours,

[6] Sean McCarthy, « How to Write a Competitive Proposal for Horizon 2020 », le 18 octobre 2016, Lyon, France.

et a eu 8 participants. Cette formation a été réalisée dans le cadre de la collaboration avec la SATT Grand Est.

2.5.6 Vulgarisation

Depuis 2013, j'interviens dans 2 modules à l'Ecole Doctorale SPIM (Sciences Pour l'Ingénieur et Microtechniques) ED 37. Ces cours sont issus de mes activités de recherche et traitent du Web de données liées et de l'informatique en nuage. Exceptionnellement, Les cours sont enseignés en anglais et en français. Il s'agit des modules :

> "Web Of Data" – 2 journées de 7h (je suis responsable de ce module)

> "Cloud Computing and Big Data" – 1 journée de 7h, en collaboration avec Prof. Christophe Nicolle).

En **octobre 2014,** j'ai soumis une proposition de MOOC (*Massive Online Open Course*) pour la plateforme FUN (France Université Numérique) traitant de l'ingénierie des connaissances. J'ai soumis cette proposition en collaboration avec d'autres collègues du département IEM de l'Université de Bourgogne et des collègues de l'équipe OUN (Objets et Usages Numériques) du laboratoire ELLIADD (Edition, Littératures, Langages, Informatique, Arts, Didactique, Discours) EA4661 de l'Université de Franche-Comté. Cette proposition a été acceptée par le CA de l'Université de Bourgogne en octobre 2015.

En **décembre 2015,** j'ai été en charge de journées de formation à des industriels bourguignons (Groupe SEB), sur l'ingénierie des connaissances. La formation a eu une durée totale de 14h, réparties sur 2 jours, et a eu 8 participants. Cette formation a été réalisée dans le cadre de la SATT Grand Est.

En **octobre 2016,** j'ai réalisé une interview pour l'association mediaConstruct (chapitre français de buildingSmart International) présentant ce qu'on entend par "Web sémantique" et quel est son intérêt pour le BIM. L'interview est disponible en ligne.

3 Recherches menées depuis le doctorat

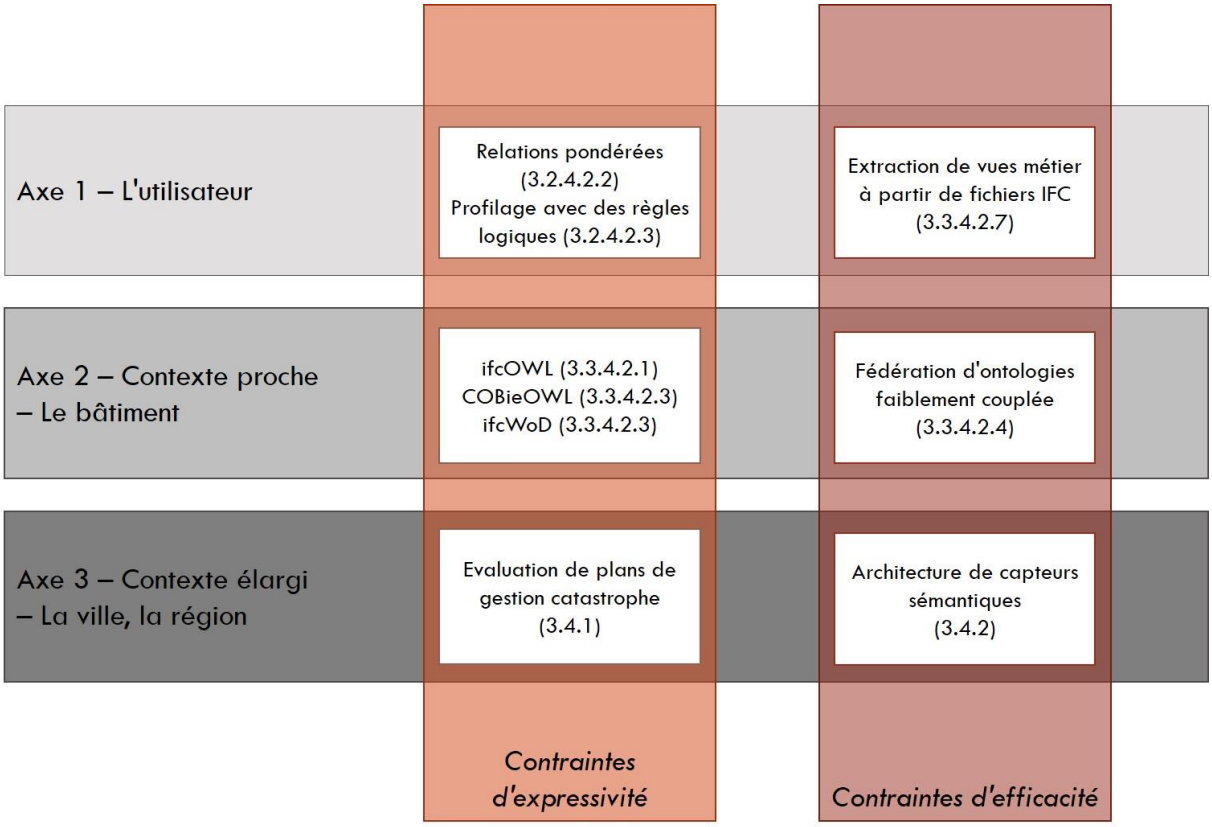
Ce chapitre présente les différents travaux de recherche réalisés depuis l'obtention de mon doctorat, plus particulièrement depuis ma nomination en tant que Maître de Conférences à l'Université de Bourgogne. Afin d'essayer de dégager un fil conducteur, j'ai choisi une présentation chronologique de mes activités. En effet, comme mes recherches concernent des domaines d'application bien différents, il est difficile de pouvoir les regrouper par thème.

Je commencerai par une présentation du contexte au sein duquel se sont déroulées mes recherches (voir section 3.1), à savoir l'ingénierie des connaissances. Je présenterai ensuite les approches et résultats obtenus au travers de deux projets distincts, correspondant chacun au co-encadrement d'une thèse CIFRE soutenue, mais dans des domaines bien différents – une thèse concernant la publicité en ligne (section 3.2), l'autre thèse traitant des maquettes numériques bâtiment (section 3.3). Sur la base des analyses, discussions et travaux futurs pour ces deux projets, je décris mon projet prospectif, qui concerne l'intégration d'approches granulaires au sein d'ontologies. Ceci représente une idée phare pas seulement pour mes recherches futures, mais que je souhaite déjà appliquer dans mes deux thèses que je co-encadre en ce moment. C'est pour cela que la section 3.4 débutera avec la présentation des différentes hypothèses derrière ce projet prospectif, puis je m'efforcerai d'illustrer l'intérêt de l'approche, ainsi que les verrous auxquels cette orientation pourrait permettre de répondre, par rapport au contexte des deux thèses que je co-encadre actuellement. La section 3.4.1 présentera donc les articulations entre ma vision recherche et la problématique à adresser dans le cadre de la thèse de Madame Claire PRUDHOMME, à savoir assurer une compréhension entre agents informatiques afin d'évaluer des plans de réponse à des catastrophes naturelles. De la même manière, la partie 3.4.2 discutera de l'apport des approches granulaires dans les systèmes de capteurs, ce qui correspond à la thèse de doctorat de Monsieur Muhammad ARSLAN. Enfin, la section 3.5 présentera les conclusions générales de ce mémoire, d'une part concernant les projets terminés et les résultats obtenus, d'autre part concernant les verrous restant à traiter et les différentes actions à mener pour y répondre.

Les travaux présentés dans ce mémoire peuvent être structurés selon 3 niveaux principaux, selon l'échelle à laquelle ils s'appliquent :

- **Axe 1 – L'utilisateur** – Les travaux dans cet axe concernent principalement la conception d'approches sensibles à un contexte (utilisateur ou d'utilisation). Cette orientation dérive de mes travaux de thèse mais aussi de ceux effectués durant mon postdoc sur les sémantiques pour l'apprentissage en ligne. En effet, j'ai travaillé sur une approche intégrant un noyau sémantique dont le but est de communiquer avec les autres modules sémantiques de la plate-forme, puis, à l'aide du langage de requêtes SPARQL, d'extraire des informations de descriptions RDF et d'inférer à partir des différents concepts définis. Ce noyau sémantique rend disponibles, pour l'apprenti, les ressources spécifiques liées à son domaine d'apprentissage, grâce aux descriptions sémantiques des modules connectés (le Blog sémantique, le Wiki sémantique, le Forum sémantique, etc.). Cette approche a été implémentée dans le cadre d'un système d'apprentissage en ligne dans le domaine de la radioprotection médicale et de la physique nucléaire.
- **Axe 2 – Le contexte proche** – Ce deuxième axe a pris naissance presque dès le début de ma thèse de doctorat, et fut fortement impacté par l'entreprise avec laquelle j'ai effectué ma thèse CIFRE (conception de solutions de géolocalisation et de diffusion de contenus selon la position de l'utilisateur) ainsi que par ma collaboration avec le Laboratoire de Topométrie de l'EPFL (description sémantique de service Web permettant de traduire une position géographique, en passant d'un système de repérage à un autre). Les problématiques de recherche sous-jacentes concernent la définition de relations spatiales (e.g. relations topologiques) et de concepts spatiaux à l'aide d'axiomes contenant des prédicats spatiaux, mais aussi les possibilités de raisonner sur la spatialité des instances (pouvoir inférer des relations spatiales). J'ai repris ces différents questionnements dans le contexte du co-encadrement des thèses des Messieurs MENDES de FARIAS et ARSLAN.
- **Axe 3 – Le contexte éloigné** e.g. le territoire – Pareillement, je poursuis des travaux dans cet axe depuis le début de ma thèse de doctorat, notamment depuis mon implication dans le projet PM2G (Plateforme Multi-Métiers Géolocalisés) dans le cadre duquel j'ai conçu l'ontologie de la viabilité hivernale dans le Territoire de Belfort, mais aussi EU FP7 ASSET où j'ai spécifié plusieurs services Web sémantiques permettant d'adapter un itinéraire par rapport à des événements imprévus (e.g. chutes de neige, accident, routes fermées, brouillard) ou encore EU FP7 TELEFOT (implémentation de l'appel d'urgence 112).

Deux axes transverses peuvent être définis, afin d'illustrer si les contraintes les plus fortes portaient sur l'expressivité du modèle ou sur l'efficacité de l'implémentation. Ceci est illustré par la figure suivante.



Plus particulièrement, par rapport aux 3 axes définis et au regard des travaux présentés dans ce mémoire, les problématiques recherche spécifiques sont :

> **Axe 1 – Sensibilité au contexte de l'utilisateur** – le verrou recherche concerne la définition de profils des utilisateurs et de leur contexte, par rapport à différents domaines métier (marketing en ligne, essais cliniques, cuisine, etc.). Ces travaux de recherche sont appliqués dans l'implémentation de systèmes de recommandation de contenus, sensibles au contexte de l'utilisateur et au contexte d'utilisation. Cette thématique de recherche représente la poursuite de mes travaux de thèse.

> **Axe 2 – Interopérabilité sémantique des Systèmes d'Information** – cet axe a pris une importance croissante ces dernières années (cela peut se voir à travers mes publications) et représente un axe fédérateur par rapport à mes recherches aujourd'hui. Le verrou de recherche concerne l'utilisation de modèles à base d'ontologies et de règles logiques afin d'intégrer et d'interconnecter des modèles hétérogènes de données. Après une étude détaillée des modèles de données sous-jacents, il s'agit de les fédérer afin de permettre de raisonner (e.g. restructuration, déduction de faits implicites) sur ces modèles et données. Il s'agit aussi de pouvoir requêter une telle fédération de modèles, et notamment optimiser le temps d'exécution de ces requêtes.

> **Axe 3 – Web sémantique spatial (interopérabilité BIM/SIG)** – il s'agit d'un axe de recherche que je poursuis depuis ma thèse CIFRE, lorsque je travaillais activement avec des SIG (ArcGIS, MapInfo) en industrie. J'ai de par le passé publié un chapitre d'ouvrage sur l'information géographique [CH9], un autre sur les communications inter-véhicules [CH10]. Les travaux de recherche associés trouvent des connexions avec les deux premiers axes exposés ici. En effet, le but visé ici est faire interopérer différents modèles et données (axe 2) afin de délivrer des réponses pertinentes par rapport à un contexte donné (axe 1). L'objectif à long terme est de délivrer une interopérabilité BIM/SIG (en suivant l'impulsion donnée par l'ISO).

Le domaine de recherche devient celui de ***l'intelligence artificielle*** (IA), cherchant à comprendre le monde à travers des modèles conceptuels et à émuler l'intelligence humaine. Les méthodes sous-jacentes, liant ces 3 thématiques, sont celles de l'ingénierie des connaissances. Il s'agit d'utiliser les ***technologies du Web sémantique*** (e.g. ontologies, données liées) et les ***logiques de description***, afin d'implémenter des systèmes pouvant ***simuler un raisonnement humain sur des données reparties sur la toile***. La section suivante définit les principaux concepts et hypothèse nécessaires pour comprendre les approches et résultats présentés dans les sections ultérieures.

3.1 Cadre de travail

Au fur et à mesure des projets, collaborations et rencontres, mes thématiques de recherche ont sensiblement évolué. Toutefois, un point commun à l'ensemble des travaux présentés dans ce mémoire est qu'ils se basent sur l'ingénierie de connaissances. Plus particulièrement je m'intéresse à la conception et l'implémentation d'approches pouvant simuler un raisonnement humain sur des données reparties sur la toile.

Pour ce faire, mes outils sont principalement les technologies du Web sémantique et des données liées. La conception d'ontologies ne peut toutefois satisfaire l'ensemble des verrous et contraintes soulevés par le déluge de données que nous connaissons actuellement.

En effet, avec des données de plus en plus reparties et hétérogènes, l'idée d'un modèle ontologique unique permettant une interprétation "fiable" de ces données par les ordinateurs semble de moins en moins pertinente. Comme mentionné plus haut, les ontologies sont des approches communément utilisées lorsqu'il s'agit de traiter l'hétérogénéité sémantique. De nombreux articles de recherche, publiés durant la dernière décennie, ont proposé le développement de nouvelles ontologies et vocabulaires pour mitiger cette hétérogénéité. Or, ce qui l'on observe c'est qu'une fois les articles publiés, les ontologies qui y sont décrites disparaissent rapidement de la toile et ne sont que très peu réutilisées par la communauté. C'est pour répondre à ces problèmes, et aussi car les ontologies ont leurs limites, que je me suis focalisée durant les dernières années sur des approches basées sur les données liées.

Ce chapitre introductif se veut un bref rappel des principes définissant le Web sémantique et les données liées.

3.1.1 Résumé du chapitre

Le principal but du Web sémantique est de se construire, au-dessus du Web actuel (ou Web syntaxique), afin de donner à chaque donnée un sens bien défini, pouvant être interprété par un ordinateur. L'idée sous-jacente de base est de construire un Web de données, réutilisant l'architecture du World Wide Web pour permettre le partage de données structurées à une échelle globale. Ceci se traduit à travers 4 principes simples, directement déduits à partir des standards sur lesquels a été construit le Web actuel.

Ces 4 principes spécifient comment identifier les ressources, comment y accéder, comment les modéliser et comment les interconnecter. Leur application pour l'ensemble des ressources (à la fois objets réels et concepts abstraits) permet d'implémenter l'idée du Web de données (ou Données liées). Le Web sémantique représente un concept plus vaste que le Web de données. Il repose sur l'utilisation de structures plus complexes pour regrouper les ressources et définir des relations et contraintes complexes entre elles, à savoir les ontologies.

Les sections suivantes présentent les concepts nécessaires à la compréhension de la nature des données et contenus du Web sémantique. Au niveau de la structuration en sous-parties, et pour faciliter la lecture du manuscrit, j'ai choisi de commencer par présenter la vision du Web de données, puis de décrire les fonctionnalités additionnelles apportées par les langages de description d'ontologies et les langages de description de règles. Je terminerai cette partie avec une section discutant des principales caractéristiques des données sémantiques par rapport aux données issues d'approches relationnelles ou orientées-objet.

3.1.2 Architecture en couches du Web sémantique

Afin d'établir le fil directeur pour ce chapitre posant les bases scientifiques des recherches présentées dans ce manuscrit, j'ai choisi de suivre l'architecture en couches du Web sémantique. Cette architecture en couches (voir Figure 1) a l'avantage de fournir une vue d'ensemble de la relation entre les différents standards utilisés pour l'identification des données, leur accès ou encore leur requêtage.

Dans la figure ci-dessous, la couche inférieure correspond à la couche identification des ressources. Elle contient les mécanismes de référencement (URI) et d'encodage des caractères (Unicode). La deuxième couche introduit le langage XML en tant que format de sérialisation standard. La troisième couche représente la couche "échange d'informations". Elle définit le modèle de données RDF (*Ressource Description Framework*) comme modèle standard pour la description des données. La couche au-dessus représente la couche des langages pouvant être utilisés pour définir des modèles ontologiques, à savoir RDF Schema, OWL (*Web Ontology Language*) et SKOS (Simple Knowledge Organisation System). Ces langages n'ont pas la même expressivité, et c'est à ce niveau que je ferais la séparation entre "données liées" et "Web sémantique" au sens large. Je vais reprendre les constatations du W3C, selon lesquelles les données liées reposent sur des vocabulaires contrôlés (définis en utilisant RDFS), alors que les applications sémantiques vont nécessiter la conception de modèles ontologiques plus complexes (définis à l'aide de la famille des langages OWL 1 et OWL 2). La couche "Requête" introduit le langage SPARQL (*SPARQL Protocol and RDF Query Language*) et comme elle concerne surtout la mouche "modèle de données", je présenterai le langage SPARQL dans la section dédiée aux données liées (3.1.3.4). La définition de règles logiques introduite par la couche "Logique" est quant à elle plus liée aux considérations sur les ontologies. Je présenterai donc le langage SWRL (utilisé jusqu'ici dans mes recherches) ainsi que les différents raisonneurs pouvant manipuler de telles règles, dans la section dédiée au Web sémantique (3.1.4.3.2).

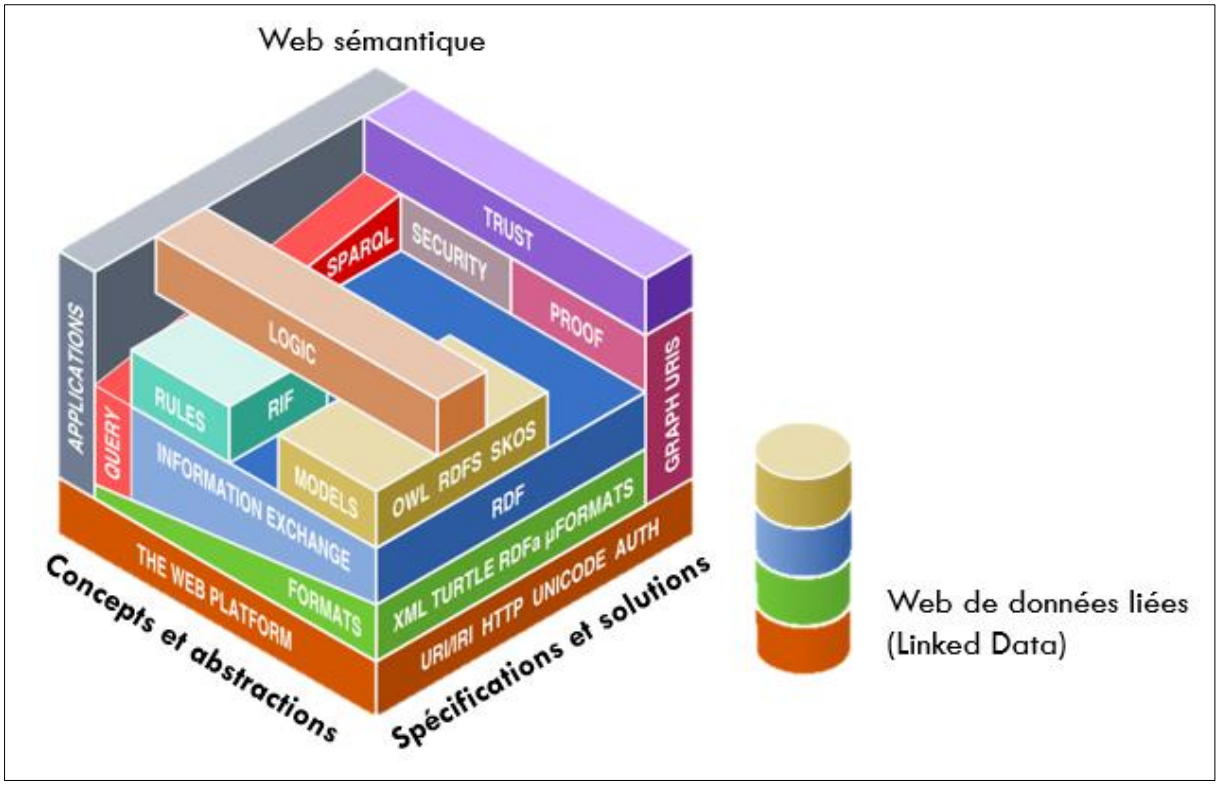


Figure 1. Architecture en couches du Web sémantique (adaptée d'après https://baojiebaojie.files.wordpress.com/2011/04/semantic_web_technology_stack.png)

Suivant l'architecture illustrée ci-dessous (Figure 1), les données liées n'utilisent pas l'ensemble des technologies du Web sémantique. Considérée par certains comme une application du Web sémantique, l'approche dite des "données liées" (ou *Linked Data*) représente un sous-ensemble des principes et des technologies du Web sémantique visant le partage et la réutilisation des données à l'échelle du Web.

Dans ce qui suit, nous verrons d'abord ces principes, puisqu'ils sont directement dérivés de l'architecture du Web actuel, mais aussi car ils correspondent aux premières couches de l'architecture ci-dessous. La section suivante présente comment le Web de données liées s'est construit au-dessus du Web actuel à travers l'identification des ressources par URIs, l'accès aux ressources via le protocole http, la modélisation de ces ressources selon le modèle RDF, avec le langage RDFS, et enfin comment requêter des ressources ainsi modélisée grâce au langage SPARQL.

Par la suite nous nous intéresserons aux modélisations formelles de la connaissance humaine, notamment à travers les ontologies définies avec la famille de langages OWL. Dans ce contexte, je vais présenter les langages de règles logiques ainsi que les raisonneurs et leurs avantages pour la manipulation et l'implémentation efficace de tels modèles.

Je clôturerai ce chapitre en résumant les caractéristiques particulières des approches sémantiques lorsque comparées aux approches plus traditionnelles de modélisation de données et connaissances.

3.1.3 Du Web syntaxique vers le Web de données liées

Le concept "données liées" n'est pas synonyme du concept "Web sémantique". En effet, les données liées reprennent une sous-partie des standards du Web sémantique, mais pas la totalité. Les principes des données liées ont été définis par Tim Berners-Lee dans (Berners-Lee 2006). Cette approche se base sur l'architecture actuelle du Web, et l'étend afin de permettre l'implémentation de raccourcis calculatoires en exploitant la structuration et l'interconnexion des différents ensemble de données implémentant ces principes. L'idée sous-jacente est de relier des silos de données isolés en un graphe global géant.

Afin de mieux comprendre les principes des données liées, nous allons faire un bref rappel des principes de l'architecture du Web actuel.

Mécanisme pour...	**Web syntaxique**	**Web des données liées**
Identifier les ressources	URI (Uniform Resource Identifier)	URI
Accès aux ressources	http (hypertext transfer protocol)	http
Présentation des ressources	HTML (Hypertext Markup Language)	RDF
Interconnexion des ressources	Hyperliens `<a href="...">`	Liens RDF

Tableau 1. Comparatif des différents mécanismes utilisés par le Web syntaxique et le Web de données

La lecture du tableau ci-dessous (Tableau 1) nous donne les 4 principes des données liées, que nous allons décrire plus en détail ci-dessous:

- Le premier principe des données liées stipule que les éléments (documents, objets réels, concepts abstraits) doivent être identifiés avec des références URI. Cela peut être vu comme une extension des principes du Web pour comprendre tout objet ou concept.
- Le deuxième principe préconise d'utiliser des URI HTTP pour accéder aux ressources. Une URI HTTP combine un mécanisme d'identification unique et un mécanisme de récupération simple. Il est ainsi possible de récupérer le contenu référencé en utilisant le protocole HTTP.
- Le troisième principe conseille d'utiliser les standards sémantiques (RDF, SPARQL) pour la publication de données structurées. Le modèle RDF (*Ressource Description Framework*) se base sur une structure en graphe, et sera détaillé plus loin dans ce chapitre.
- Le quatrième principe est le plus important dans la vision des données liées – il conseille d'inclure, dans les descriptions sémantiques des éléments, des liens vers les URI d'autres éléments, afin de permettre leur découverte. Le quatrième principe propose d'utiliser les liens RDF afin de connecter toutes sortes d'éléments et pas seulement des documents.

3.1.3.1 L'identification par URI

Ce principe stipule que l'identification des ressources dans le Web de données liées se fait en utilisant le même mécanisme d'identification qu'utilisé actuellement dans le Web. Il s'agit des URI (*Uniform Ressource*

Identifier). Chacun des trois mots a son importance. Une URI est un identifiant de ressources (pas nécessairement quelque chose d'accessible via Internet), elle permet donc la résolution d'identité à l'échelle du Web. Le caractère "uniforme" permet :

> Différents types d'identifiants de ressources peuvent être utilisés dans un même contexte, même si les mécanismes d'accès aux ressources diffèrent
> Une interprétation sémantique uniforme des différentes conventions syntaxiques utilisées pour les identifiants de ressources
> De nouveaux types d'identifiants de ressources sans interférer avec ceux déjà existants

Une URI est une chaîne de caractères identifiant une ressource sur un réseau, qu'elle soit abstraite ou physique. Tous les hyperliens sont exprimés sous forme d'URI. Une URI doit respecter la syntaxe générique définie par la norme RFC3986 (T. Berners-Lee 2005):

```
URI = schema"://"[userinfo"@"]host[:port][path]["?"query]["#"fragment]
```

Figure 2. Syntaxe générique des URI adaptée selon (T. Berners-Lee 2005)

Terme	Explication
schema	syntaxe associée à un protocole de la couche "Application" du modèle OSI (e.g. http, ftp, mailto, telnet, irc, ssh, etc.)
userinfo	e.g. username:password
host	nom de domaine, adresse IPv4/IPv6
port	e.g. 80 pour le port http standard
path	chemin d'accès dans le système de fichiers du serveur Web
query	paramètres de requête
fragment	identifie une partie (fragment) d'un document

Tableau 2. Syntaxe générique des URI.

Dernièrement, a été définie une extension de la syntaxe générique des URI, afin de supporter l'encodage Unicode / ISO 10646. Grâce à la norme RFC3987[7] il est possible de définir un mapping entre URI et IRI. Il est dès lors possible d'utiliser des IRIs à la place des URIs pour identifier des ressources.

```
http://با.ثت.حج/د خ/رذ/سز؟صش=ضظ;عظ=فغ#كق
```

Figure 3. Exemple d'une IRI.

Les URIs fonctionnent bien pour exprimer des identités sur le Web, mais deviennent assez pénibles à manipuler lorsque l'on souhaite les inclure dans des modèles faits "sur papier". Il est possible d'utiliser un schéma d'abréviation d'URI, appelé *qname*. Dans sa forme la plus simple, une URI exprimée selon ce schéma comprend 2 parties : un nom de domaine et un identifiant, séparées par ":". Par exemple, si on considère le nom de domaine **geo**, la représentation au format *qname* pour le concept "**Angleterre**" serait **geo:Angleterre**. Il est important de remarquer le fait que les *qnames* ne sont pas des identifiants globaux (ne s'appliquent pas à l'échelle du web, mais à l'intérieur du nom de domaine considéré). Seules les URI qualifiées intégralement représentent des identifiants Web globaux. Dès lors, chaque représentation d'un *qname* doit être accompagnée par une définition de la correspondance du nom de domaine.

3.1.3.2 L'accès par URI http

Le second principe stipule que, dans le Web de données, l'accès aux ressources se fera sur la base du même mécanisme d'accès universel qu'utilise aujourd'hui le Web. Il s'agit du protocole HTTP (Hyper Text

[7] http://www.ietf.org/rfc/rfc3987.txt

Transfert Protocol) (R. Fielding 1999). En effet, le Web est un espace de données, pouvant être utilisé à la fois par les humains et par les machines. Les deux parties doivent être capables de récupérer des représentations des ressources dans une forme qui leur convient ; il s'agit de représentations HTML pour les humains, et de descriptions RDF pour les machines. Cela peut être réalisé grâce au mécanisme de négociation du contenu du protocole HTTP (R. Fielding 1999).

Toutefois dans un contexte de données liées, il existe 2 principales stratégies pour créer des URI identifiant des ressources déréférençables : URI 303 et URI avec ancre. Les 2 stratégies assurent que les objets et les documents qui les décrivent ne sont pas confondus et que les humains autant que les machines peuvent récupérer des formats appropriées. La note du groupe d'intérêt "Cool URIs for the Semantic Web[8]" du W3C les spécifie.

3.1.3.3 La modélisation avec RDF

Nous venons de mentionner le fait que pour que les machines puissent manipuler des descriptions de ressources, celles-ci doivent être au format RDF. Ceci représente le troisième principe des données liées. En effet, suivant l'exemple du langage HTML qui permet de définir des pages Web, il faut utiliser le modèle de données RDF (Hayes & Patel-Schneider 2014) pour la description de ressources.

3.1.3.3.1 Des données tabulaires vers les données graphe

Historiquement, les données sont généralement représentées sous forme de tableaux (données tabulaires), dans lesquels chaque ligne représente un élément que nous décrivons, et chaque colonne représente les propriétés de cet élément. Les cellules (intersection d'une ligne avec une colonne) correspondent à des valeurs spécifiques des propriétés considérées.

Lorsque l'on considère le problème de la répartition de ces données sur le Web (une partie des données sera sur un ordinateur, tandis que d'autres parties seront sur d'autres ordinateurs), plusieurs approches sont possibles:

> Première approche - **Distribution par ligne** - Dans cette approche, chaque machine connectée au réseau maintien des informations concernant une ou plusieurs lignes (complètes) du tableau ci-dessus. Une requête concernant un élément donné ne peut avoir de réponse que de la part de l'ordinateur stockant les informations sur la ligne correspondante. Cette solution offre suffisamment de flexibilité, puisque les machines peuvent partager la charge induite par la représentation de plusieurs éléments individuels. Mais vu qu'il s'agit d'une représentation répartie des données, elle nécessite une sorte de coordination entre les serveurs. Notamment, chaque serveur doit partager les informations concernant les colonnes (e.g. pour savoir si la 2e colonne sur un serveur correspond à la même information que la 2nde colonne sur un autre serveur). Ce problème n'est pas insurmontable, et il est même fondamental pour la répartition/distribution des données.

> Deuxième approche - **Répartition par colonne** - Dans cette approche, chaque serveur est responsable d'une ou plusieurs colonnes (complètes) du tableau initial. Dans cet exemple, un serveur est responsable de publier les types d'œuvre et les dates, un autre serveur étant responsable des titres. Cette solution est aussi flexible, mais d'une manière différente que l'était la solution précédente. En effet, cette approche permet à chaque machine d'être responsable d'un seul type de données. Si on n'est pas intéressé par les années de publication, les données contenues sur ce serveur ne sont pas nécessaires. Il est aussi possible de spécifier de nouvelles informations concernant les différents éléments (e.g. combien de pages fait l'œuvre considérée) sans influer sur les informations stockées sur les autres serveurs. Cette approche nécessite aussi la mise en place d'une procédure de coordination des identifiants des éléments à décrire. Encore une fois, la coordination est nécessaire afin de déterminer si la ligne 3 sur un serveur fait bien référence au même élément que la ligne 3 sur un autre serveur.

> Troisième approche – **RDF ou combinaison des 2 approches précédentes** - Dans ce cas, les données ne sont pas réparties ligne par ligne ou colonne par colonne, mais cellule par cellule. Chaque machine est responsable d'une ou plusieurs cellules. Ce type d'approche combine la flexibilité des 2 approches précédentes. 2 serveurs peuvent s'échanger la description d'un élément et ils peuvent

[8] https://www.w3.org/TR/cooluris/

partager l'usage d'une propriété donnée. Ce niveau de flexibilité est nécessaire si l'on souhaite respecter le slogan AAA "*Anyone can say Anything about Any topic*" (devise du Web). Cette approche permet de formuler un énoncé concernant un élément donné et de spécifier n'importe quelle propriété de l'élément considéré – elle combine donc les avantages des deux premières approches. Elle en combine aussi les coûts - non seulement, une référence globale pour les en-têtes des colonnes est nécessaire, mais il faut aussi référencer les lignes. En fait, la représentation de chaque cellule se fait par le biais de 3 valeurs : une référence globale pour sa ligne, une référence globale pour sa colonne et la valeur contenue dans la cellule considérée.

C'est cette dernière stratégie qui représente le parti pris par RDF.

3.1.3.3.2 Le modèle de données - RDF

Puisque dans l'approche considérée, une cellule est représentée via trois valeurs, l'unité de base en RDF est appelée un triplet.

> L'identifiant de la ligne est appelé sujet du triplet.
> L'identifiant de la colonne est appelé prédicat du triplet (puisque les colonnes spécifient des propriétés des entités contenues dans les lignes).
> La valeur de la cellule est appelée objet du triple.

Le modèle de données RDF représente l'information comme un graphe dirigé avec des nœuds et des arcs nommés. Dans un contexte "données liées" :

> Le sujet d'un triplet est l'URI identifiant la ressource décrite.
> L'objet peut être soit une valeur littérale (une chaîne de caractères, un nombre ou une date), soit une URI d'une autre ressource qui est liée, d'une manière ou d'une autre, à l'objet.
> Le prédicat, au milieu, indique le type de relation qui relie le sujet à l'objet (par exemple, le nom ou le numéro de sécurité sociale, pour une valeur littérale ; une connaissance ou un membre de la famille, pour une autre ressource). Il est, lui aussi, identifié par une URI. Ces URI de prédicats proviennent des "vocabulaires". Ce sont des collections d'URI qui peuvent être utilisées pour représenter l'information dans un certain domaine (par exemple FOAF[9]).

La recommandation RDF (Hayes & Patel-Schneider 2014) contient de nombreuses fonctionnalités. Toutefois, dans un contexte de données liées, plusieurs de ces fonctionnalités sont déconseillées (e.g. la réification, les collections, les conteneurs et les nœuds anonymes). En effet, la vision du Web de données est plus réduite que celle du Web sémantique, où effectivement l'ensemble des fonctionnalités RDF peuvent être utilisées. La restriction sur un sous-ensemble du modèle RDF est surtout là pour faciliter la vie des consommateurs de données.

Le Tableau 3 illustre plusieurs énoncés RDF. Le premier énoncé définit l'auteur du document dont l'URI est "https://cv.archives-ouvertes.fr/ana-roxin". Les énoncés suivants définissent respectivement la liste de publications et le nom associés à la ressource identifiée par l'URL "https://orcid.org/0000-0001-9841-0494".

Enoncé 1	Sujet	https://cv.archives-ouvertes.fr/ana-roxin	L'URL https://cv....
	Prédicat	http://purl.org/dc/elements/1.1/creator	… a pour auteur …
	Objet	https://orcid.org/0000-0001-9841-0494	La ressource identifiée par l'URL https://orcid.org...
Enoncé 2	Sujet	https://orcid.org/0000-0001-9841-0494	La ressource identifiée par l'URL https://orcid.org...
	Prédicat	http://xmlns.com/foaf/spec#publications	… a pour liste de publications …
	Objet	https://cv.archives-ouvertes.fr/ana-roxin	L'URL https://cv....
Enoncé 3	Sujet	https://orcid.org/0000-0001-9841-0494	La ressource identifiée par l'URL https://orcid.org...
	Prédicat	http://xmlns.com/foaf/spec#name	… a pour nom …
	Objet	"Ana Roxin"	"Ana Roxin"

Tableau 3. Exemples d'énoncés RDF.

[9] Friend Of A Friend - http://xmlns.com/foaf/spec/

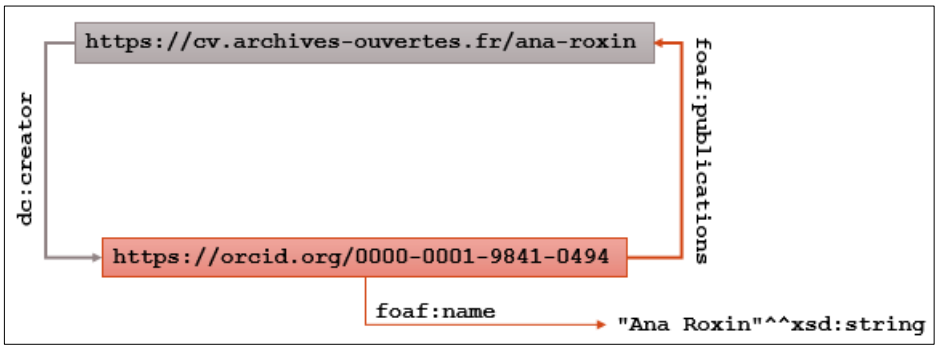


Figure 4. Représentation sous forme de graphe des énoncés du Tableau 3.

Enfin, les ressources étant modélisées avec RDF, il faut utiliser le protocole SPARQL pour les recherches ou les requêtes de ressources.

3.1.3.3.3 Le langage de description de vocabulaires pour RDF - RDFS

Comme mentionné plus haut, RDF n'est pas un format de données, mais représente un modèle de données simple pour décrire un ensemble ("graphe orienté") de ressources interconnectées par des relations (appelées prédicats ou propriétés). RDF ne fournit aucun mécanisme pour déclarer de telles ressources et propriétés. Ceci est le rôle du langage RDF Schema (ou RDFS). En RDF, le sujet et l'objet d'un triplet (ou d'un énoncé) peuvent être n'importe quoi. Toutefois, en modélisation de données, il est important de pouvoir faire la différence entre des types de "n'importe quoi":

- Premier type: quelque chose qui a une représentation directe dans la réalité, une instance précise dans le temps
  - CETTE table est une instance / MA Voiture est une instance du concept VOITURE
- Deuxième type: quelque chose de plus conceptuel, qui potentiellement regroupe plusieurs "n'importe quoi" du premier type.
  - UNE table est un concept

Afin de permettre cette différenciation, RDFS introduit un nouveau vocabulaire au-dessus de RDF, en fournissant ainsi les moyens de définir des termes qui seront utilisés dans des énoncés RDF tout en leur associant un sens précis.

Classes RDFS	**Propriétés RDFS**
`rdfs:Class` définit un objet abstrait appliqué (via rdf:type) pour créer des instances	**`rdfs:subClassOf`** propriété transitive définissant des hiérarchies de classes
`rdf:Property`	**`rdfs:subPropertyOf`** propriété transitive pour définir des hiérarchies de propriétés
`rdfs:Literal`	**`rdfs:domain`** définit le domaine d'une propriété (en termes de classes)
`rdfs:Resource` toute entité d'un modèle RDF est une instance de cette classe	**`rdfs:range`** définit la portée d'une propriété (en termes de classes)
`rdfs:Datatype`, **`rdf:XMLLiteral`**, **`rdfs:Container`**, **`rdfs:ContainerMembershipProperty`**	

RDFS constitue un modèle permettant de définir des vocabulaires simples, regroupant plusieurs énoncés RDF (**`rdf:Statement`**). RDFS permet de définir les propriétés d'une ressource mais aussi les types de ressources étant décrits ainsi.

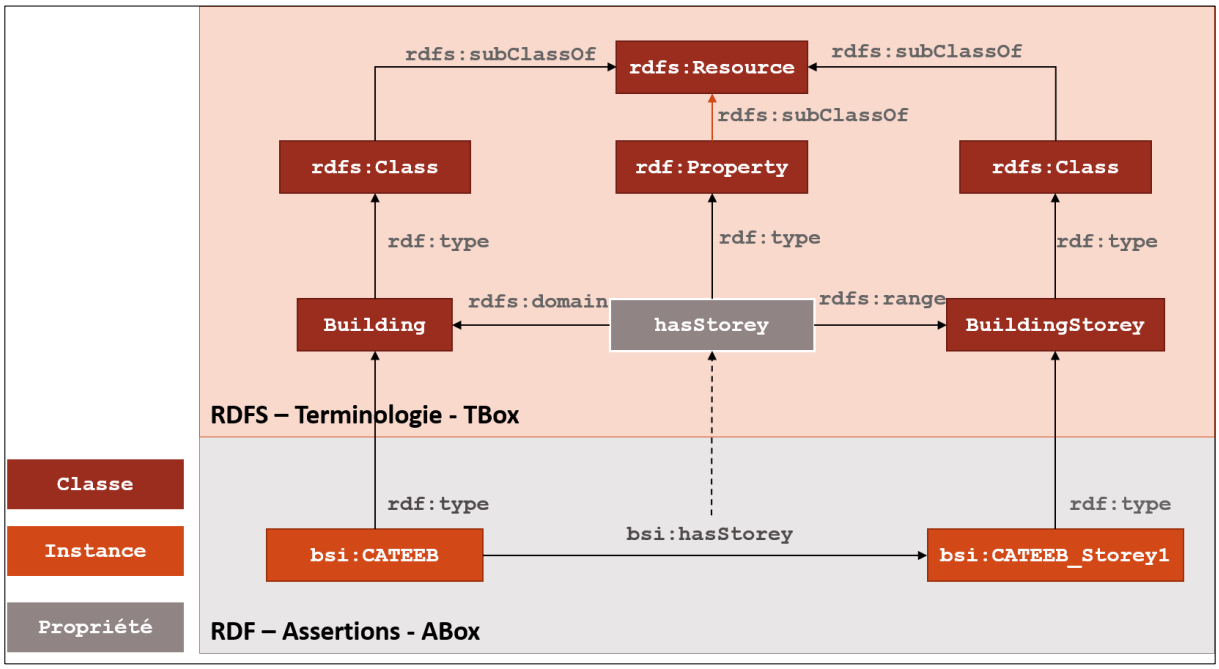


Figure 5. Illustration des nouveaux concepts apportés par le langage RDFS par rapport à RDF.

La sémantique d'un terme d'un vocabulaire RDFS est donnée à travers ses propriétés et ses valeurs. Les figures suivantes (Figure 6, Figure 7 et Figure 8) illustrent les déductions (ou *inférences*) qu'il est possible d'effectuer dans un vocabulaire RDFS.

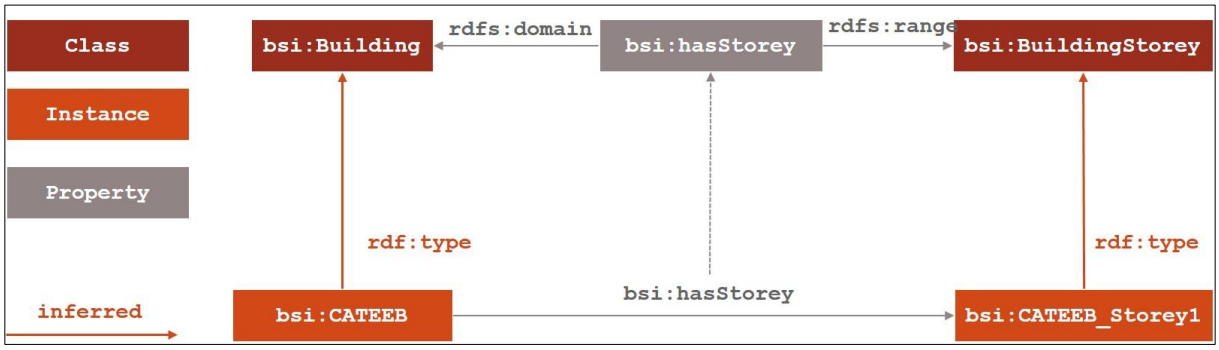


Figure 6. Déduction de l'appartenance d'une instance à une classe à partir du domaine et/ou de la portée d'une propriété

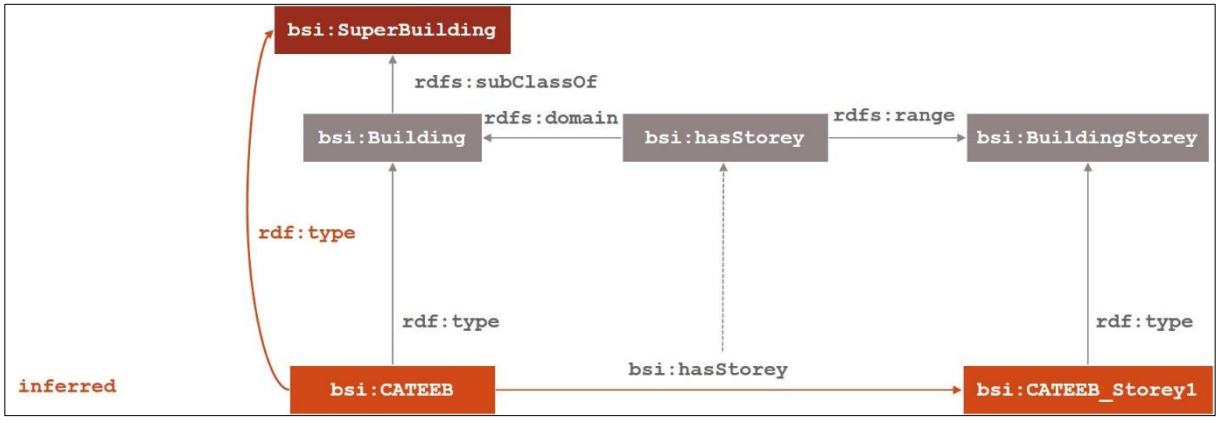


Figure 7. Déduction de l'appartenance à une superclasse à partir d'une hiérarchie de classes

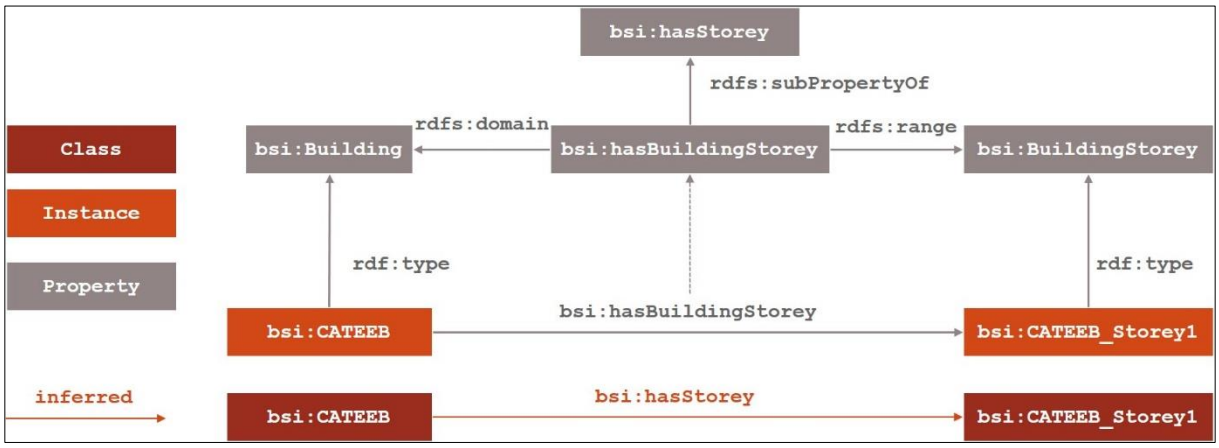


Figure 8. Déduction de nouveaux triplets à partir d'une hiérarchie de propriétés

3.1.3.4 Le langage de requêtes SPARQL

SPARQL (*SPARQL Protocol and RDF Query Language*) représente un ensemble de protocoles et langages permettant de requêter des données stockées dans des bases de connaissances. Publié dans sa version 1.0 en janvier 2008 par le W3C, SPARQL en est aujourd'hui à sa version 1.1 (Harris & Seaborne 2013), version qui a été reconnue standard W3C depuis mars 2013.

Dans sa version 1.0 (Prud'hommeaux & Seaborne 2008), SPARQL permet l'extraction de données sous la forme soit de nœuds blancs ou valeurs littérales (typées ou pas), soit en tant que sous-graphes RDF. Il supporte aussi l'exploration de données, la requête de relations inconnues, l'exécution de jointures sur des bases hétérogènes en une seule requête, la transformation de données RDF d'un vocabulaire à un autre, et enfin la construction de nouveaux graphes RDF avec RDF Query Graphs. La version 1.0 de SPARQL comprend uniquement les 3 spécifications suivantes (mises à jour avec SPARQL 1.1):

- Langage de requête pour des graphes RDF (*SPARQL Query Language Specification*) - http://www.w3.org/TR/2013/REC-sparql11-query-20130321/
- Protocole permettant l'adressage de requêtes via http (*SPARQL Protocol for RDF Specification*) - http://www.w3.org/TR/sparql11-overview/#sparql11-protocol
- Format XML pour le rendu des résultats aux requêtes (*SPARQL Query XML Results Format*) - http://www.w3.org/TR/2013/REC-rdf-sparql-XMLres-20130321/

Figure 9. Structure de base d'une requête SPARQL.

SPARQL est basé sur la sérialisation Turtle de RDF et sur un appariement de modèles de graphes. Un modèle de graphe (ou "*graph pattern*" en anglais) est un triplet RDF qui contient des variables à un emplacement arbitraire, soit en tant que Sujet, soit en tant que Prédicat, soit en tant qu'Objet. Dans la figure ci-dessous, le modèle de graphe est illustré en tant que modèle de triplet.

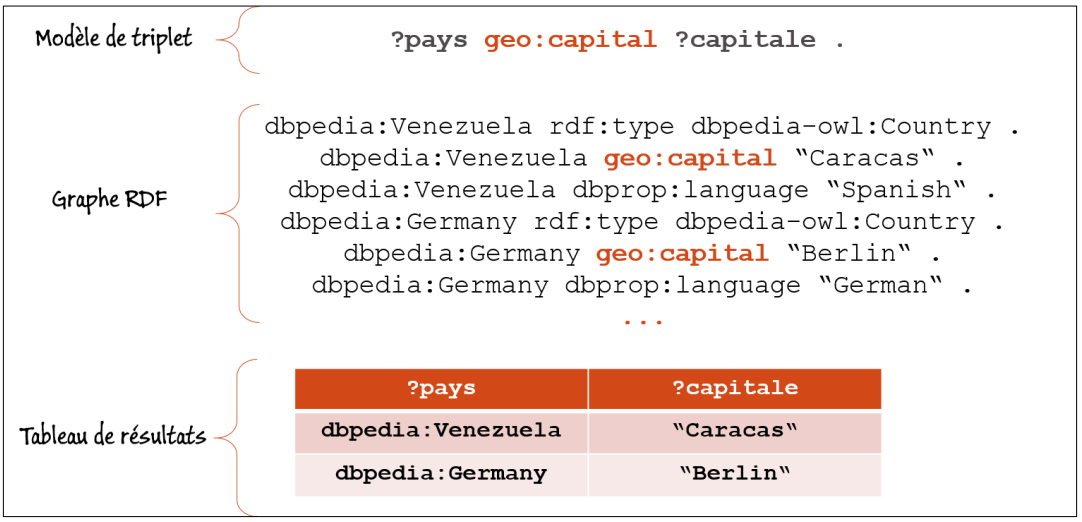


Figure 10. Exemple d'interprétation du modèle de triplet.

Dans sa version 1.1 (Harris & Seaborne 2013), SPARQL a été enrichi avec des fonctionnalités supplémentaires e.g. fonctions d'agrégation, sous-requêtes, prise en charge des négations, possibilité de construire des chaînes de propriétés (ou "*property path*"" en anglais). Un support des implications logiques (ou "*logical entailment*" en anglais) a été ajouté pour RDF(S), OWL Direct et RIF (*Rule Interchange Format*) Core. Parmi les principales nouveautés, on peut citer SPARQL Update, un langage complet pour la manipulation (mise à jour) des données de graphes RDF, et la possibilité d'exécuter des requêtes fédérées (ou "*federated querries*" en anglais) sur des points d'accès SPARQL répartis. En plus des spécifications ci-dessus, les spécifications suivantes ont été ajoutées:

- Spécification Federation Extensions – comment une seule requête peut retourner des résultats à partir de sources multiples
- Spécification Update – permet l'ajout, la modification ou la suppression de données
- Spécification Service Description – comment un programme client peut demander à un moteur SPARQL les fonctionnalités qu'il supporte
- Query Results JSON Format – équivalent JSON du Query Result XML Format
- Graph Store HTTP Protocol – étend le protocole SPARQL avec une API type REST pour la communication entre un client et un moteur SPARQL
- Spécification Entailement Engines – quelles informations prendre en compte lors des inférences
- SPARQL New Features and Rationale – liste des nouvelles fonctionnalités SPARQL 1.1

3.1.3.5 L'interconnexion entre ressources

Nous avons décrit RDF comme une approche permettant de répartir les données sur plusieurs sources. Toutefois, lorsque l'on souhaite réutiliser ces données, il faut réassembler ces différentes sources. Un des avantages des représentations à base de triplets est de faciliter l'assemblage des données ainsi réparties.

Lorsque l'on souhaite assembler plusieurs graphes, il est essentiel de pouvoir déterminer quand un nœud d'un graphe est identique à un autre nœud dans un autre graphe. Le modèle RDF résout ce problème en utilisant les URI. La solution implémentée par le modèle RDF vient du modèle traditionnel du Web, à savoir le référencement à base d'URIs. La syntaxe et le format d'une URI sont familiers, même aux utilisateurs novices du Web, et ceci à cause du cas particulier des URL. Toutefois, la signification d'une URI en tant qu'identifiant global pour une ressource Web n'est pas appréciée à sa juste valeur. Si deux agents Web souhaitent référencer une même ressource, la pratique recommandée est d'utiliser une même URI pour identifier cette ressource – cette pratique n'est pas spécifique au Web sémantique, mais s'applique au Web en général.

Pour simplifier, une URI est un identifiant à but global (à échelle du Web). Dès lors, quelles que soient 2 applications Web elles pourront faire référence à la même ressource en référençant la même URI. C'est la syntaxe d'une URI qui rend son déréférencement possible, en d'autres termes qui permet d'utiliser l'ensemble des informations contenues dans l'URI (e.g. nom du serveur, protocole, numéro de port, nom

de fichier, etc.) pour localiser un fichier (ou un emplacement dans un fichier) sur le Web. D'un point de vue modélisation, c'est la possibilité de déréférencer une URI qui permet aux différents modèles ainsi conçus de prendre part à la construction d'une infrastructure Web globale (*Giant Global Graph*).

Toujours en suivant l'exemple du Web qui utilise des hyperliens pour connecter des données présentes dans différents serveurs, la définition de liens RDF entre ressources (ou des jeux de données) permet de constituer un espace d'information global. Les liens RDF permettent de modéliser des relations entre 2 ressources, et comprennent 3 références d'URI (2 pour le sujet et l'objet qui identifient les ressources liées, et une URI pour le prédicat qui définit le type de relation entre les ressources) – ce sont des propriétés objet dans le contexte d'une ontologie.

> Les liens RDF internes relient des ressources dans une seule source de données liées (les URI des sujets et des objets sont dans le même espace de noms) ;
> Les liens externes connectent des ressources dans des sources de données liées différentes (URI des sujets et objets des liens externes sont dans des espaces de noms différents).

Il existe 3 principaux types de liens RDF:
> Les **liens de relation** pointent vers des éléments dans d'autres sources de données. Par exemple ajouter des données bibliographiques sur les publications.
> Les **liens d'identité** pointent vers des alias d'URI (URI différentes mais se référant à la même ressource) utilisés par d'autres sources de données pour identifier le même objet réel ou concept abstrait. Des liens d'identité permettent aux clients de récupérer d'autres descriptions d'une entité à partir d'autres sources de données. Ces liens ont une fonction sociale importante, puisqu'ils permettent d'exprimer des vues différentes du monde sur le Web de données.
> Les **liens de vocabulaire** spécifient des liens allant des données vers les termes de vocabulaire qui les représentent et de ces définitions vers les définitions de termes apparentés dans d'autres vocabulaires. Ils rendent les données auto descriptives, ce qui fait que les applications de données liées comprennent des données de plusieurs vocabulaires et les intègrent. Les langages OWL (*Web Ontology Language*, langage d'ontologies pour le Web), RDFS (*RDF Schema*, schéma RDF) et SKOS (*Simple Knowledge Organization System*, système simple d'organisation des données) définissent des types de liens RDF qui peuvent être choisis à cette fin (voir Tableau 4).

Terme à utiliser pour définir un lien RDF	Description
`owl:equivalentClass, owl:equivalentProperty`	Utilisés pour déclarer que les termes de différents vocabulaires sont équivalents.
`rdfs:subClassOf, rdfs:subPropertyOf`	Utilisés pour déclarer une correspondance moins forte
`skos:broadMatch, skos:narrowMatch`	

Tableau 4. Exemples de prédicats pouvant être utilisés pour définir des liens de vocabulaire.

3.1.4 Web sémantique

Le Web Sémantique peut être vu comme une extension du Web de données manipulant non plus de simples données interconnectées mais des connaissances liées. La principale différence va résider dans la complexité et l'expressivité des modèles utilisés pour représenter la connaissance. Alors que typiquement les ressources manipulées en tant que données liées auront des représentations sémantiques relativement simples ou peu complexes, manipuler des connaissances implique de travailler avec des modèles plus complexes, à savoir les ontologies.

En effet, alors qu'RDFS est bien adapté à la modélisation des vocabulaires simples, il a des limites qui n'en font pas un candidat adapté pour la modélisation de connaissances plus complexes. Si l'on suppose le modèle décrit ci-dessous (voir Figure 11), à savoir un bâtiment a des étages qui peuvent être de deux types: soit "sous-sol", soit "étage". En l'état, la recommandation RDFS ne nous permet pas de spécifier qu'un bâtiment sans sous-sol ne doit pas contenir d'étage de type sous-sol. Les ontologies et les langages permettant de les définir viennent combler ces manques.

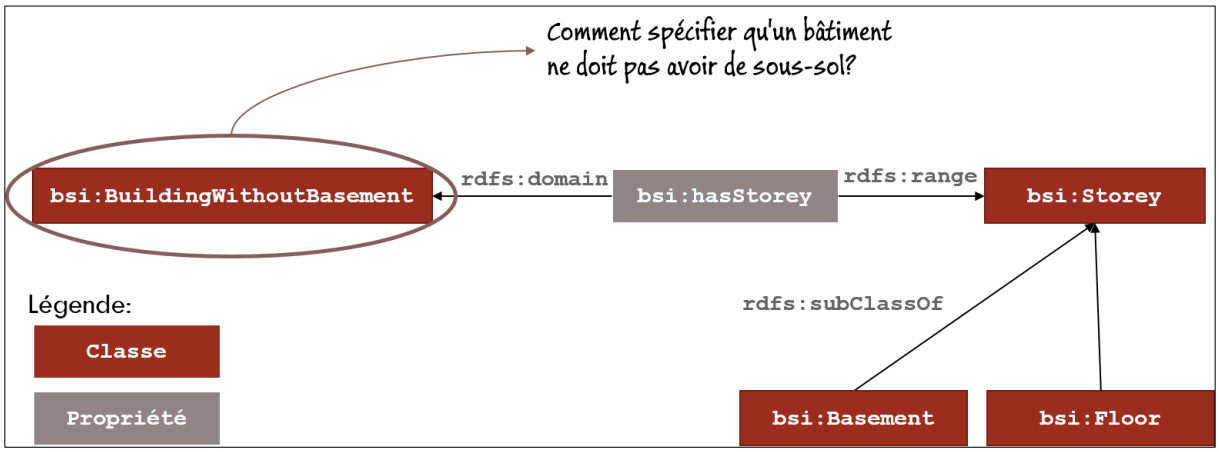


Figure 11. Illustration des limites de RDFS.

Les ontologies apparaissent comme les éléments de base du Web sémantique. Les ontologies sont des modèles conceptuels capturant et rendant explicite le vocabulaire utilisé dans des applications sémantiques. Les ontologies garantissent une communication sans ambiguïtés. Elles sont considérées en tant que *lingua franca* du Web sémantique.

3.1.4.1 Les ontologies en tant que support pour le raisonnement

3.1.4.1.1 Les ontologies

Afin de dépasser les limites de RDFS, il est possible d'utiliser la famille de langages OWL pour définir des ontologies plus expressives. La figure suivante illustre les différents types de représentations de la connaissance pouvant être définis, ordonnés selon leur niveau de formalisme (les modèles formels étant les plus expressifs). Selon cette organisation (Figure 12), les ontologies se situent à l'extrémité droite de la flèche - ce sont des modèles formels spécifiés à l'aide de langages logiques.

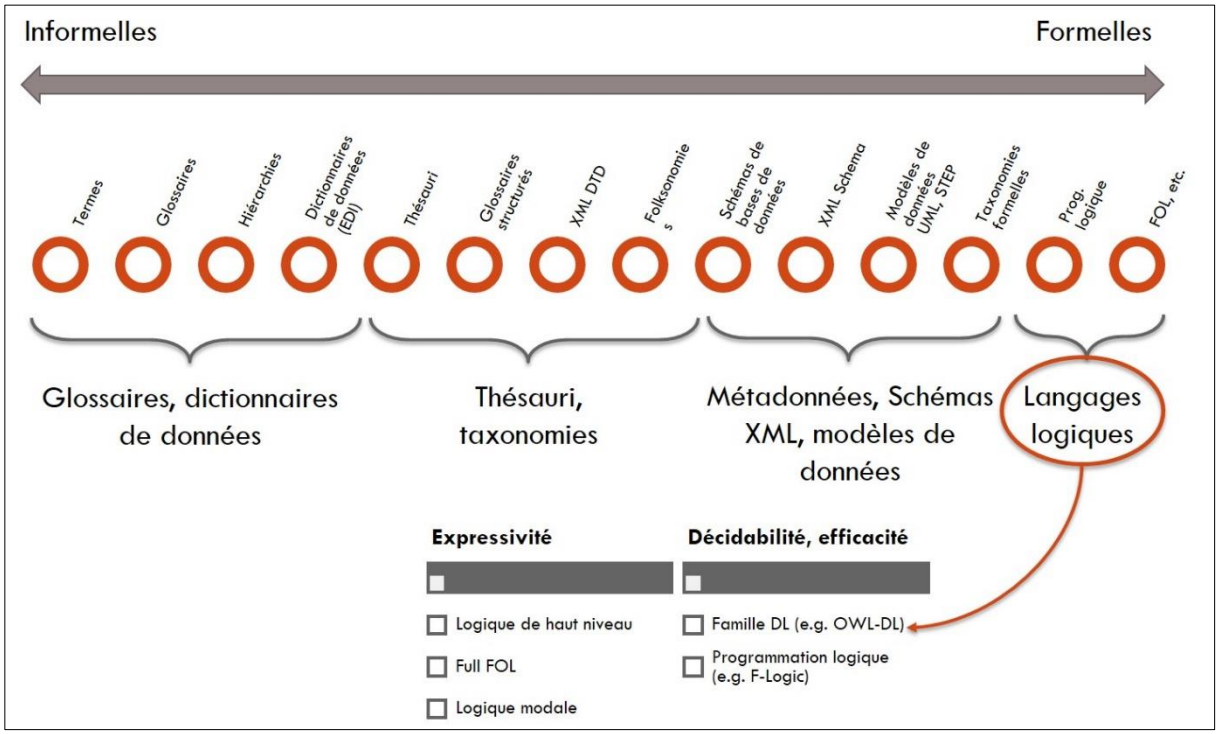


Figure 12. Différents types de modèles – des plus informels vers les plus formels.

Les ontologies étant des représentations formelles, elles sont décrites en utilisant des langages logiques. Selon la priorité à donner dans la modélisation des connaissances, à savoir favoriser l'expressivité ou la décidabilité, plusieurs choix de langages de modélisation d'ontologies peuvent être faits. Si l'on souhaite

maximiser l'expressivité du modèle, il est possible d'utiliser de langages logiques de haut niveau (e.g. Full FOL). Si au contraire, on souhaite augmenter l'efficacité d'implémentation, il est possible d'employer des langages basés sur la logique de description (e.g. OWL Full) ou sur la programmation logique (e.g. F-Logic).

Définition 1. *(Ontologie au sens large)* Une ontologie est une "spécification formelle et explicite d'une conceptualisation partagée d'un domaine de connaissance" (Studer et al. 1998)

Dans cette définition, chaque terme est important:

- **Conceptualisation** - Il s'agit d'une vue abstraite, simplifiée du monde que nous souhaitons représenter avec un but quelconque.
- **Spécification** - Une conceptualisation peut être spécifiée à travers des structures relationnelles extensionnelles (ou langages formels) ou intentionnelles (langages informels). Toutefois une ontologie représente une conceptualisation dans un domaine précis. Il n'est donc pas nécessaire de considérer l'ensemble des mondes possibles comme c'est le cas dans les structures relationnelles intentionnelles (représentées à travers des triples $\mathbf{C} = (\boldsymbol{D}, \boldsymbol{W}, \boldsymbol{R})$, où $\boldsymbol{W}$ est l'ensemble des mondes possibles, et $\boldsymbol{R}$ un ensemble de relations conceptuelles dans l'espace domaine $< \boldsymbol{D}, \boldsymbol{W} >$). Par rapport à un domaine de connaissance précis, il suffit de considérer le domaine du discours ($\boldsymbol{D}$) et l'ensemble des relations permises dans ce domaine ($\boldsymbol{R}$).
- **Formelle** - Les expressions doivent pouvoir être manipulées par les ordinateurs
- **Explicite** - La spécification de la conceptualisation doit contenir le sens de tous les concepts qu'elle contient
- **Partagée** - On ne spécifie pas la connaissance d'une personne, mais un consensus général de la réalité.

Les définitions suivantes, extraites à partir de (Grimm et al. 2011), sont plus "pragmatiques" et précisent d'autres composantes des ontologies, notamment en termes d'entités manipulées et niveau d'expressivité.

Définition 2. *(Ontologie)* Une ontologie $\boldsymbol{O}$ représente un tuple $\boldsymbol{O} = (\boldsymbol{S}, \boldsymbol{A})$, $\boldsymbol{S}$ représentant la signature de l'ontologie (les éléments conceptuels utilisés pour représenter la connaissance) et $\boldsymbol{A}$ étant l'ensemble d'axiomes associé (exprimé dans un langage de description d'ontologie ou un autre formalisme de représentation de la connaissance).

Définition 3. *(Signature d'une ontologie)* La signature d'une ontologie est définie en tant que l'union $\boldsymbol{S} = \mathbf{C} \cup \mathbf{I} \cup \mathbf{P}$ de l'ensemble des classes $\mathbf{C}$, des instances $\mathbf{I}$ et des propriétés $\mathbf{P}$.

Définition 4. *(Entités conceptuelles et entités concrètes)* Si l'on distingue les objets abstraits du domaine considéré et les valeurs concrètes des données, les ensembles $\mathbf{C}$, $\mathbf{I}$ et $\mathbf{P}$ deviennent $\mathbf{C} = \boldsymbol{C} \cup \boldsymbol{D}$ (concepts et types de données $\mathbf{I} = \boldsymbol{I} \cup \boldsymbol{V}$ (instances et valeurs), $\mathbf{P} = \boldsymbol{R} \cup \boldsymbol{T}$ (relations $\mathcal{R}$ et types $\mathcal{T}$).

Une ontologie est décrite en utilisant des langages de description d'ontologies et la spécification résultante contient le schéma de connaissance pour l'ensemble des classes $\mathbf{C}$ et des propriétés $\mathbf{P}$ considérées.

Comme mentionné plus haut, une ontologie représente une "axiomatisation » des connaissances. En effet, à l'intérieur d'une ontologie, les connaissances sont exprimées à travers des axiomes de subsumption, domaine, portée ou encore disjonction. Le Tableau 5 liste les axiomes les plus utilisés et donne les expressions associées en logique du premier ordre (Grimm et al. 2011). Pour l'appariement de ces axiomes par rapport aux langages de description d'ontologies existants, le lecteur peut consulter (Sowa 1999).

Type d'axiome	Définition	Notation	Expression en logique du premier ordre
Instanciation	Affecte une instance à une classe	$\alpha_{\wedge}(i, C)$	$[(C, i)], i \in \mathrm{I}, c \in \mathrm{C}$
Assertion	Définit une relation entre deux instances par le biais d'une propriété	$\alpha_{\rightarrow}(i_1, \mathrm{p}, i_2)$	$[\mathrm{p}(i_1, i_2)], i_1, i_2 \in \mathrm{I}, p \in \mathrm{P}$
Subsumption	Pour deux classes, spécifie que toute instance de la classe subsumée est aussi une instance de la classe subsumable	$\alpha_{\triangle}(E_1, E_2)$	$[\forall x: E_1(x) \rightarrow E_2(x)], E_1, E_2 \in \mathrm{C} \cup \mathrm{P}$
Domaine	Définit la classe représentant le domaine d'une propriété	$\alpha_{\mathrm{D}\rightarrow}(p, D)$	$[\forall x, y: p(x, y) \rightarrow D(x)], p \in \mathrm{P}, D \in \mathrm{C}$
Portée	Définit la classe représentant la portée d'une propriété	$\alpha_{\rightarrow R}(p, R)$	$[\forall x, y: p(x, y) \rightarrow R(x), p \in \mathrm{P}, R \in \mathrm{C}$
Disjonction	Spécifie que l'intersection de deux classes est l'ensemble vide (elles n'ont aucune instance en commun)	$\boldsymbol{\alpha_{\oplus}(C_1, C_2)}$	$[\forall \boldsymbol{x}: \boldsymbol{C_1}(\boldsymbol{x}) \wedge \boldsymbol{C_2}(\boldsymbol{x}) \perp], \boldsymbol{C_1}, \boldsymbol{C_2} \in \mathbf{C}$

Tableau 5. Types d'axiomes communément utilisés dans les formalismes d'ontologies - d'après (Domingue et al. 2011)

3.1.4.1.2 Les langages de description d'ontologies

Ayant maintenant défini les ontologies et les éléments manipulés, il convient de discuter du formalisme logique que j'ai adopté dans mes recherche pour la représentation des ontologies. Comme mentionné plus haut, il est possible d'utiliser la logique du premier ordre ou les langages basés sur la logique de description. Les paragraphes suivants représentent une justification du choix des langages à base de logique de description dans mes travaux.

Dérivée de la logique propositionnelle (LP), la logique du premier ordre ou FOL (*First Order Logic*) comprend des éléments supplémentaires à savoir les quantificateurs existentiel et universel, les opérateurs de la logique des propositions, des variables, des constantes, des fonctions et des relations ou prédicats. La syntaxe associée définit les règles pour la formulation correcte de:

- Termes à partir de Variables, Constantes et Fonctions
  - f(X), g(a,f(Y)), s(a), calories(Plat)
- Atomes à partir de Relations avec Termes en arguments
  - p(f(X)), q(s(a), g(a,f(Y))), ajouter(a, s(a), s(a)), superieur_a(calories(Plat), 500)
- Formules à partir d'Atomes, Opérateurs et Quantifieurs
  - (∀Plat) (superieur_a(calories(Plat), 500)) ⟶ calorique(Plat))

Grâce aux règles de syntaxe et aux opérateurs supportés, la logique du premier ordre permet de modéliser une sémantique par rapport à une structure composée d'un domaine $\boldsymbol{D}$ ainsi que de symboles relations, fonctions et constantes mappés en tant que respectivement relations sur $\boldsymbol{D}$, fonctions dans $\boldsymbol{D}$ et éléments de $\boldsymbol{D}$. Ainsi les assertions deviennent des éléments du domaine $\boldsymbol{D}$, les relations avec des arguments peuvent être évaluées à **VRAI** ou **FAUX**, et il est possible de combiner tout ceci avec des opérateurs logiques ainsi que les quantifieurs existentiel et universel. A titre d'exemple, le tableau ci-dessous (Tableau 6) présente plusieurs assertions et donne leur modélisation en logique du premier ordre.

Assertions complexes	Modélisation
Tous les enfants aiment la glace.	∀X: Enfant(X) ⟶ aimeGlace(X)
Il existe des plats végétariens qui sont bons.	∃X: PlatVegetarien(X) ∧ Bon(X)
La mère d'un veau est une vache.	∀X, ∀Y: Veau(Y) ∧ estMere(X,Y) ⟶ Vache(X)
La relation r est symétrique	∀X, ∀Y: r(X,Y) ⟶ r(Y,X)

Tableau 6. Exemples d'assertions en logique du premier ordre.

La logique du premier ordre supporte aussi les déductions:

> Une théorie T est un ensemble de formules (propositions)
> Une interprétation I est un modèle de T si et seulement si I ⊨ G pour toutes les formules G de T.
> Une formule F est une conséquence logique de T si et seulement si tous les modèles de T sont aussi des modèles de F. On écrit alors : T ⊨ F
> Deux formules F et G sont dites équivalentes d'un point de vue logique si et seulement si {F} ⊨ G et {G} ⊨ F

La logique du premier ordre permet donc la définition de systèmes déductifs. Un tel système est dit correct, si les déductions effectuées respectent la sémantique du système. Un tel système est dit complet, si la véracité des formules valides sur le plan sémantique peut être aussi démontrée au niveau syntaxique (Loveland 1978).

L'implémentation de processus de déduction en logique du premier ordre repose sur l'utilisation de procédures de décision, à savoir des algorithmes vérifiant les 3 propriétés suivantes:

> **Arrêt** - L'algorithme doit donner le résultat en un temps fini
> **Correction** - Les inférences produites sont en accord avec la sémantique associée. Ce qui est vrai sur le plan syntaxique l'est sur le plan sémantique.
> **Complétude** - Toutes les formules valides (e.g. vraies sur le plan sémantique) peuvent être démontrées sur le plan syntaxique.

Dérivé de la logique des propositions, l'algorithme Tableaux demeure une des procédures de décision les plus utilisées de nos jours pour la satisfiabilité de concepts. Inventé à la base en tant que méthode Tableaux sémantique par Evert Willem Beth en 1955, il a été davantage simplifié et adapté en méthode Tableaux analytique par Raymond Merrill Smullyan dans son ouvrage "First Order Logic" (Smullyan 1968). Conçue initialement pour la logique des propositions, la méthode Tableaux est basée sur une preuve par réfutation (la cohérence d'une formule logique est vérifiée en inférant que sa négation est une contradiction). Lorsque comparée à la logique propositionnelle, la logique du premier ordre possède en plus le quantifieur universel ainsi que le quantifieur existentiel. L'algorithme Tableaux comporte donc des règles additionnelles pour traiter les formules contenant des quantifieurs. Son adaptation pour la logique du premier ordre fut aussi faite par Smullyan dans le même ouvrage (Smullyan 1968).

Revenant à l'hypothèse de départ selon laquelle nous pourrions utiliser la logique du premier ordre pour la modélisation d'ontologies, oui, il est possible d'utiliser la logique du premier ordre pour décrire des ontologies. Malheureusement, cela reviendrait à programmer des applications Web en langage assembleur ! Autant il est possible d'utiliser la logique du premier ordre pour les ontologies, autant cela n'est pas du tout adapté pour les raisons suivantes:

> La logique du premier ordre est bien trop expressive et trop volumineuse pour la modélisation d'ontologies
> Elle ne facilite pas non plus la recherche de consensus entre choix de modélisation
> Elle n'est que semi-décidable et encore des preuves théoriques très complexes doivent être mises en œuvre

Toutefois, ce ne sont pas ces raisons qui en font un candidat à éliminer ! Il s'agit de trouver un fragment de la logique du premier ordre qui soit plus approprié pour la description d'ontologies, c'est-à-dire un fragment de FOL qui soit moins expressif et décidable. En l'occurrence, ce fragment existe et il s'agit de la logique de description (DL ou *description logics* en anglais). La logique de description représente un ensemble de formalismes logiques conçus pour la représentation des connaissances, notamment à travers des graphes sémantiques. Les langages composant cet ensemble sont, la plupart du temps, décidables et suffisamment expressifs (contrairement aux algorithmes en FOL qui ne se terminent pas toujours). Comme en logique de description, il est possible de composer des descriptions plus complexes à partir de descriptions simples, en utilisant des constructeurs, les différents langages composant la famille des logiques de description se différencient par rapport aux constructeurs appliqués. C'est ce qui détermine leur niveau d'expressivité.

Le Tableau 7 présente l'ensemble des constructeurs utilisés en logique de description avec la syntaxe et la famille de langages associés.

Constructeur / opérateur	Syntaxe	Langage	
Conjonction	A ⊓ B	$\mathcal{FL}$	$\mathcal{S}$*
Restriction de valeur	∀R.C		
Quantifieur existentiel	∃R		
Tout	⊤	$\mathcal{AL}$*	
Rien	⊥		
Négation	¬A		
Disjonction	A ⊔ B		
Restriction existentielle	∃R.C		
Restriction de cardinalité	(≤nR) (≥nR)		
Ensemble d'individus	$\{a_1,...,a_2\}$		
Hiérarchie de propriétés	R ⊑ S	$\mathcal{H}$	
Propriétés inverses	R^{-1}	$\mathcal{I}$	
Restriction de cardinalité qualifiée	(≤nR.C) (≥nR.C)	$\mathcal{Q}$	

Tableau 7. Constructeurs utilisés en logique de description.

Les langages basés sur la logique de description n'ont donc pas tous le même niveau d'expressivité associé. Selon les fonctionnalités supportées, chaque langage a son propre niveau d'expressivité (voir Tableau 8). Un langage comme OWL 2 DL aura une expressivité de type $\mathcal{SHROIQ(D)}$ puisqu'il supporte le langage ALC et la définition de propriétés transitives, la définition de hiérarchies de propriétés, les chaînes de propriétés, les ensembles d'instances (prédicat `owl:oneOf`), les propriétés inverses ainsi que la définition des restrictions de cardinalités qualifiées le tout par rapport à différents types de données.

Caractéristiques	Syntaxe
Langage attributs / compléments	$\mathcal{ALC}$
ALC + propriétés transitives	$\mathcal{S}$
Hiérarchies de propriétés	$\mathcal{H}$
Ensembles d'individus (un de)	$\mathcal{O}$
Propriétés inverses	$\mathcal{I}$
Restriction de cardinalité	$\mathcal{N}$
Restriction de cardinalité qualifiée	$\mathcal{Q}$
Types de données	$\mathcal{(D)}$
Propriétés fonctionnelles	$\mathcal{F}$
Conjonction de propriétés	$\mathcal{R}$

Tableau 8. Expressivité des langages DL

Les connaissances ainsi modélisées sont stockées dans ce qu'on appelle une base de connaissances, qui les classe en connaissances terminologique et connaissances assertionnelles. Dans la suite de ce manuscrit, j'utiliserai en tant que synonymes les termes d' "ontologie[10]" et "base de connaissances". La Figure 13 précise les différents éléments contenus à l'intérieur de la TBox et respectivement de la ABox.

Définition 5. *(TBox et ABox)* L'ensemble des classes **C**, relations ***R*** et des contraintes définies sur ces classes et relations forme l'ensemble des connaissances terminologiques d'une ontologie ou TBox (*Terminological Box*). L'ensemble des instances du modèle terminologique forme les connaissances assertionnelles ou ABox (*Assertional Box*) (Haarslev et al. 2008).

[10] Comme le précise le W3C (https://www.w3.org/standards/semanticweb/ontology), il n'y a pas non plus de différence claire entre le concept d' « ontologie » et le concept de « vocabulaire », si ce n'est que la tendance générale est d'utiliser « ontologie » pour des modèles de connaissance qui sont plus complexes et surtout formels par rapport aux spécifications plus faiblement couplées et nécessitant bien moins de formalisme (et dans ce cas on utilise plutôt le terme de « vocabulaire »).

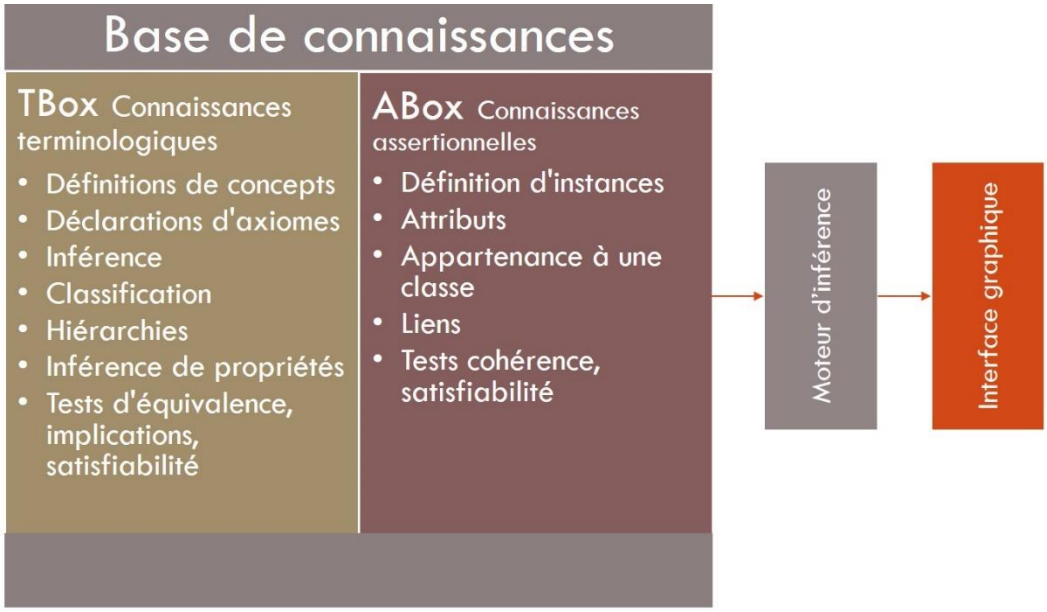


Figure 13. Architecture générale des applications basées sur la logique de description

C'est au-dessus d'une telle base de connaissances qu'un algorithme type procédure de décision (ou *raisonneur*) sera exécuté. La sortie d'un tel algorithme matérialise dans la base de connaissances l'ensemble des connaissances non spécifiées de manière explicite et donc inférées. C'est pour cela qu'on appelle un tel algorithme moteur d'inférence. Enfin lors d'une implémentation d'une approche reposant sur une telle base de connaissances, une interface graphique est souvent ajoutée afin de faciliter l'interaction avec la base de connaissances et le moteur d'inférence.

Au niveau des éléments manipulés par les ontologies, elles comprennent toujours des classes, propriétés et des instances de classes (ou des individus). En ingénierie d'ontologies, afin de préciser les conditions que doit remplir un individu pour appartenir à une classe, celle-ci est décrite de façon formelle, en spécifiant: a) des conditions nécessaires ou b) des conditions nécessaires et suffisantes. Ceci est expliqué dans la figure suivante (voir Figure 14).

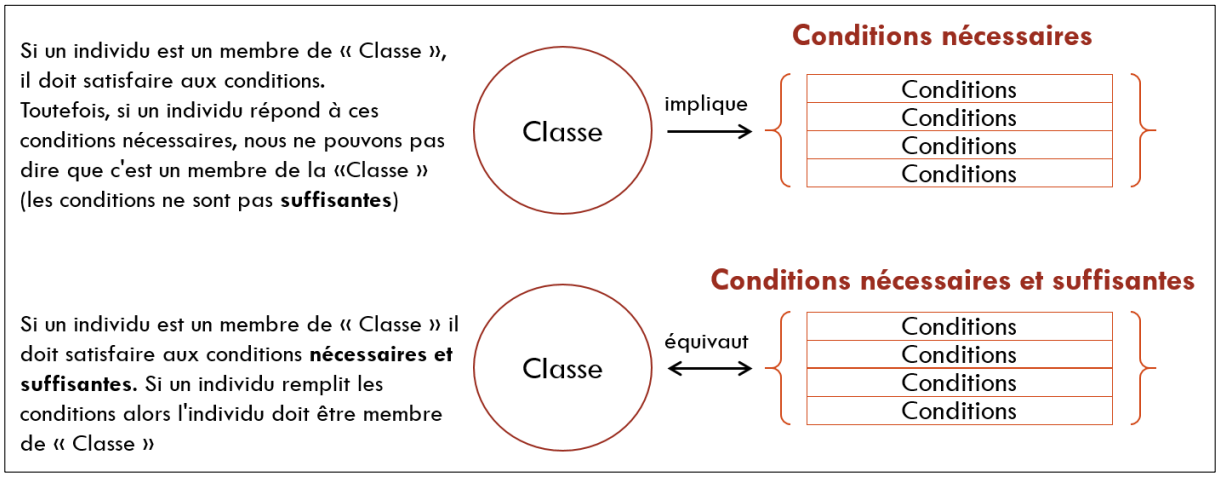


Figure 14. Différences entre conditions nécessaires et conditions nécessaires et suffisantes.

3.1.4.1.3 La famille OWL

Dans l'ensemble des recherches présentées ici, j'ai mis un point d'honneur à concevoir des modèles ontologiques décidables. Il est en effet important d'obtenir une réponse à une requête en un temps fini si on souhaite une implémentation efficace des modèles conçus. J'ai donc principalement travaillé avec les langages de description d'ontologies basés sur la logique de description (Baader et al. 2003), notamment la famille de langages OWL 1(Dean et al. 2006) et OWL 2 (Motik et al. 2012). En effet, le W3C a défini deux familles de recommandations:

- OWL 1 est une recommandation W3C depuis 2004 et qui supporte un niveau d'expressivité $\mathcal{SHOIN}(\mathcal{D})$. OWL1 représente un fragment de la Logique du Premier Ordre (FOL) et comprend plusieurs langages OWL Lite ⊆ OWL1 DL ⊆ OWL1 Full (voir Figure 15)
- OWL 2 qui est une recommandation W3C depuis 2009 et qui est plus expressive, atteignant un niveau $\mathcal{SHROIQ}(\mathcal{D})$. OWL2 est aussi un fragment de la Logique du Premier Ordre (FOL) et comprend aussi 3 langages, avec en plus 3 profils différents pour OWL2 DL, à savoir: OWL2 EL, OWL2 RL, OWL2 QL ⊆ OWL2 DL ⊆ OWL2 Full (voir Figure 16) :
  - OWL2 EL - Permet les axiomes de sous-classes avec intersection, quantificateur existentiel, classes tout, rien et fermées avec un seul membre. Ne permet pas la négation, la disjonction, le quantificateur universel, les propriétés inverses.
  - OWL2 QL - Permet les sous-propriétés, la définition de sous-classes et de domaines/portées. Ne permet pas les classes fermées.
  - OWL2 RL - Permet tous les types d'axiomes, les restrictions de cardinalités (sur la portée seulement ≤1 et ≤0), et les classes fermées avec un seul membre. Ne permet pas certains constructeurs (quantificateur universel et négation sur le domaine, quantificateur existentiel et union de classes pour la portée).

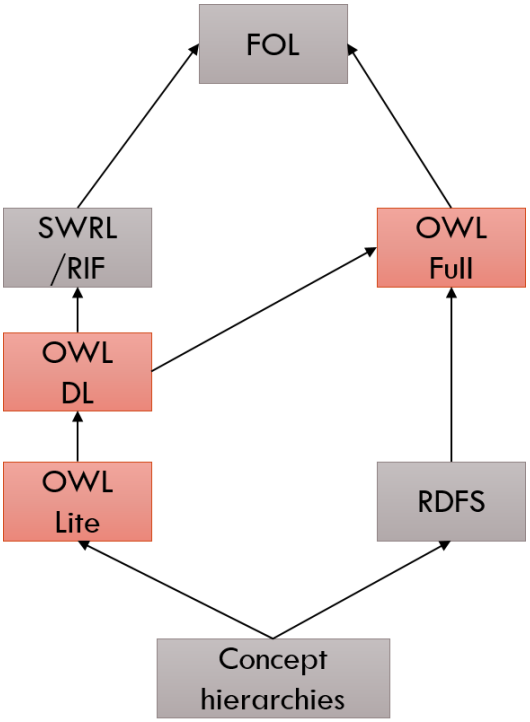


Figure 15. Famille de langages OWL1

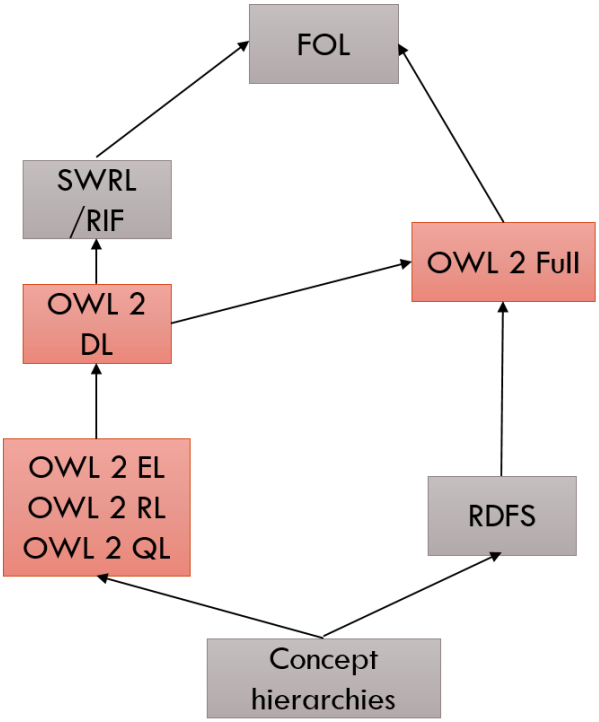


Figure 16. Famille de langages OWL2

Il existe 2 classes prédéfinies dans OWL: **`owl:Thing`** (contient tous les individus) T et **`owl:Nothing`** (la classe vide) ⊥. Il est possible de préciser que plusieurs classes sont disjointes entre elles (leur intersection est l'ensemble vide).

La famille de langages OWL supporte deux types de propriétés.

- Les propriétés dites d'objet sont définies comme des instances de la classe **`owl:ObjectProperty`**. Elles ont un ensemble de départ (domaine modélisé à travers le prédicat **`rdfs:domain`**) et un ensemble d'arrivée (portée ou **`rdfs:range`**).
- Les propriétés dites de type de données ont pour ensemble d'arrivée des types de données. Elles sont définies dans l'ontologie en tant qu'instance de la classe **`owl:DatatypeProperty`**.

La famille de langages OWL supporte aussi les hiérarchies de propriétés (se définissent toujours avec **`rdfs:subClassOf`**) ainsi que les propriétés inverses (se définissent avec le prédicat **`owl:inverseOf`**). OWL permet de définir des propriétés disjointes à travers le prédicat **`owl:propertyDisjointWith`**.

Si deux propriétés R et S sont disjointes, alors il n'y aura jamais deux individus qui seront reliés par ces propriétés. En suivant la même logique que pour la définition des concepts "tout" et "rien", OWL 2 ajoute les propriétés universelle et vide :

- **`owl:topObjectProperty`** relie deux individus quelconques dans l'univers
- **`owl:topDataProperty`** relie chaque individu avec chaque valeur
- **`owl:bottomObjectProperty`**, **`owl:bottomDataProperty`** relations vides

La figure suivante illustre les relations entre les langages de description d'ontologies à base de sémantiques classiques et les langages utilisant des sémantiques à prédicats logiques.

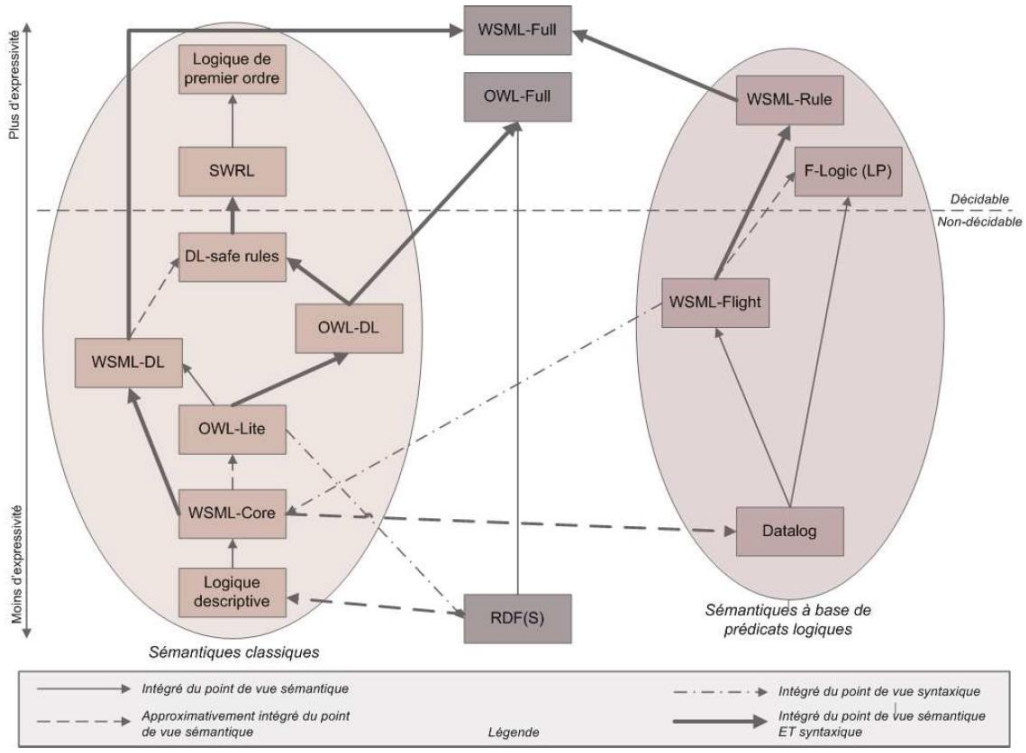


Figure 17. Relations entre les différents langages de description d'ontologies (Roxin 2009)

3.1.4.2 Le raisonnement en logique de description

Comme précisé plus haut, un raisonneur est un algorithme permettant de matérialiser des connaissances non explicites présentes dans une base de connaissances, et ce, en un temps fini. Lorsqu'on s'intéresse aux algorithmes dits "de raisonnement" en logique de description, ceux-ci peuvent être répartis en 2 familles (Napoli 1997):

- Algorithmes NC (Normalisation-comparaison)
  - Création de formes normales des concepts de la base de connaissance
  - Comparaison des formes normales en utilisant des règles de comparaison
- Algorithme des Tableaux sémantiques pour la logique de description et ses optimisations
  - S'applique aux langages de la famille DL qui intègrent la négation des concepts définis
  - "Est-ce que D subsume C?" devient "Est-ce que D ne subsume pas C?"

Comme le précise Amedeo Napoli dans son rapport de recherche INRIA (Napoli 1997), le problème avec les algorithmes NC est qu'en dehors des langages de type FL-, ils ne permettent pas de détecter l'ensemble des relations de subsomption valides. Ayant pour ma part principalement travaillé avec des ontologies modélisées avec des langages de la famille OWL, ce type d'algorithme ne me permettait pas d'obtenir les inférences souhaitées. J'ai donc travaillé avec des algorithmes de raisonnement basés sur la méthode des Tableaux sémantiques, et c'est ces approches que je vais brièvement décrire dans ce qui suit.

Nous avons mentionné l'algorithme Tableaux utilisé en logique des propositions et son adaptation pour la logique du premier ordre. Nous pourrions penser à utiliser l'algorithme Tableaux tel que défini pour

FOL, puisque la logique de description est bien un fragment de la logique du premier ordre. Malheureusement, lorsqu'utilisés en logique de description, les algorithmes FOL ne se terminent pas toujours. Cela comprend donc la décidabilité des ontologies ainsi définies. Il faut donc considérer des algorithmes avec une durée d'exécution finie et adaptés pour la logique de description. Par exemple, si l'on considère à nouveau l'algorithme Tableaux, cela revient à non plus identifier si un théorème (au sens logique du premier ordre) est respecté mais bien si une base de connaissances est respectée. Il faut adapter les aspects déductifs de l'algorithme à la détection de contradictions dans la base de connaissances. Le Tableau 9 présente des exemples d'adaptation de contradictions à identifier dans une base de connaissances (notée "KB" pour "*knowledge base*" en anglais dans le tableau ci-dessous).

Problème	**Exemple**	**Requête DL**
(In)cohérence de la base de connaissances	Est-ce que la base de connaissances est sensée ?	KB $\models \perp$?
(In)cohérence des classes	Est-ce que la classe C doit être vide ?	C $\equiv \perp$?
Subsomption de classes La connaissance est-elle correcte (intuitions capturées) ?	Structure de la base de connaissances Est-ce que C est subsumé par D selon T ?	C $\sqsubseteq$ D ? $C^I \subseteq D^I$ pour tout modèle I de T
Equivalence de classes Est-ce que la redondance de la connaissance est minime (pas de synonymes inattendus) ?	Est-ce que deux classes sont identiques ? Est-ce que C est équivalent à D selon T ?	C $\equiv$ D ? $C^I = D^I$ dans tout modèle I de T
Disjonction de classes	Est-ce que deux classes sont disjointes ?	C $\sqcap$ D $= \perp$?
Appartenance à une classe Connaissances requêtables	Est-ce que l'individu a appartient à la classe C ?	C(a) ?
Génération d'instance	Trouver toutes les instances (connues) de la classe C	

Tableau 9. Principaux tests d'inférence.

Si l'on reprend l'exemple de l'algorithme Tableaux en FOL, il s'agit d'un algorithme de preuve qui vérifie la cohérence d'une formule logique (ou d'un théorème) en inférant que sa négation est une contradiction. C'est ce que l'on appelle la preuve par réfutation. Lorsqu'appliqué en FOL, l'algorithme Tableaux revient à construire un arbre en appliquant successivement les règles d'extension de l'algorithme (les conjonctions de formules deviennent un chemin de la racine vers une feuille, une branche sur ce même chemin représentant une disjonction). Un chemin dans Tableaux est dit fermé si pour une même formule F (qu'elle soit atomique ou pas) on trouve sur le même chemin F et non F. Dans ce cas-là, le théorème de départ est considéré faux (ou la formule logique n'est pas cohérente).

Pour adapter l'algorithme des Tableaux à la logique de description, il faut transformer les formules composant la base de connaissance en forme négative normale (NNF) en appliquant les règles dites "De Morgan" autorisant uniquement les opérateurs de conjonction ($\wedge$), disjonction ($\vee$) et négation ($\neg$). Dès lors, dans une base de connaissances W, il faut l'adapter dans sa forme négative normale NNF (W). La Figure 18 illustre une application de l'algorithme Tableaux pour vérifier la formule $(q \wedge r) \vee (p \wedge \neg r) \vee r$.

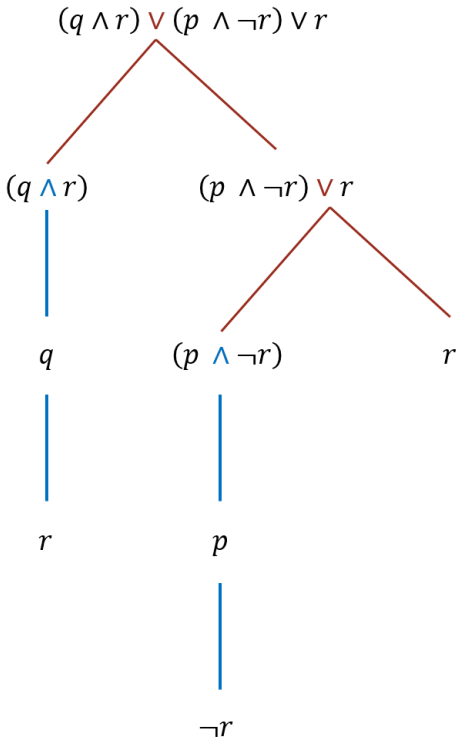


Figure 18. Exemple d'application de l'algorithme Tableaux pour prouver la validité d'une formule.

Toutefois ce n'est pas parce que l'algorithme semble simple, qu'une implémentation naïve suffit. En effet, sans faire attention il est très facile d'aboutir sur des situations de non-terminaison. Considérons le cas d'un arbre ayant seulement 10 inclusions de concepts réparties sur 10 nœuds racine (ou assertions). L'algorithme des Tableaux permet d'obtenir 2^100 expansions possibles !

En effet, la recherche d'un raisonneur "efficace" est un des "héritages" des logiques de description sur les ontologies. Un tel raisonneur est avant tout un algorithme qui se termine en un temps fini et qui permet de déduire des conclusions ou d'obtenir des résultats. Un tel algorithme doit traiter de manière informatique l'ensemble des axiomes présents dans la base de connaissances afin de déduire des connaissances implicites. Un tel processus peut être implémenté grâce à l'utilisation de langages formels, car eux seuls permettent de spécifier (et ce à travers des sémantiques formelles) les conséquences d'un ensemble d'axiomes. La plupart des raisonneurs utilisés aujourd'hui pour les ontologies représentent des optimisations de raisonneurs utilisés en logique de description e.g. Pellet (Sirin et al. 2007) ou HermiT (Glimm et al. 2014).

Parmi ces optimisations, on retrouve:

> Classification optimisée (calcul d'un ordre partiel entre concepts)
> Diffusion de règles accélérée (explorer d'information de tests précédents)
> Utiliser l'information de structure pour définir l'ordre de classification
> Tests satisfiabilité/subsumption optimisés
> Normalisation et simplification de concepts
> Absorption (simplification) des axiomes
> Dépendance de retours (*backtracking*) directs
> Sauvegarde du résultat de satisfiabilité et des modèles partiels
> Ordonnancement heuristique des extensions propositionnelles et modales

Je ne présenterai pas ici une étude exhaustive des différents raisonneurs existants pour les langages à base de logique de description. En effet, je me suis uniquement intéressée aux raisonneurs pouvant manipuler des bases de connaissances définies avec des langages de la famille OWL1 et OWL2. De plus, ayant utilisé des règles logiques (exprimées avec le langage SWRL) au-dessus des bases de connaissances, les raisonneurs utilisés devaient comprendre à la fois les axiomes OWL et les axiomes SWRL. La section suivante présente donc le langage SWRL ainsi que les différents raisonneurs pouvant manipuler à la fois des axiomes OWL et SWRL.

3.1.4.3 Raisonner sur des ontologies comportant des règles logiques

3.1.4.3.1 La définition de règles logiques

Dans la recherche concernant les formalismes de représentation de la connaissance, une tendance actuelle est d'intégrer à des ontologies employant des logiques descriptives (*DL-style ontologies*) des règles de programmation logique (*LP-style rules*). Lorsqu'on ajoute de telles règles logiques à des ontologies basées sur des langages DL, on obtient ce qu'on appelle des systèmes à base de règles. Ces systèmes représentent une des principales approches aujourd'hui dans l'implémentation de processus de raisonnement au-dessus de bases de connaissances.

Définition 6. ***(Inférence)*** On appelle inférence le mécanisme permettant de dériver de nouvelles assertions à partir des axiomes et des règles logiques existantes dans la base de connaissances ((w3c) 2015).

Définition 7. ***(Règle ou axiome de règle)*** Une règle est composée d'une tête ou conséquent (en anglais "*rule head*") et d'un corps ou antécédent (en anglais "*rule body*"). Si le corps de la règle est évalué à **`VRAI`**, alors sa tête est dérivée en tant que nouvelle assertion de la base de connaissances (Horrocks et al. 2004)(B. Motik et al. 2005).

Les systèmes à base de règles reposent sur ces règles logiques afin d'exprimer des connaissances sous la forme de paradigmes logiques du type **`SI... ALORS...`**. Ceci est considéré comme une forme spéciale d'axiomatisation (Grimm et al. 2011).

Définition 8. *(Clause de Horn)* Une clause de Horn est une implication depuis un antécédent (un ensemble de formules atomiques) vers un conséquent (une formule atomique) (Horrocks et al. 2004)(B. Motik et al. 2005)

Plusieurs langages existent et permettent d'ajouter de telles règles dans une base de connaissances, comme par exemple le langage *DL-safe rules* (Motik & Sattler 2006).

3.1.4.3.2 Le langage SWRL

Une des tentatives dans ce domaine est le langage SWRL (*Semantic Web Rule Language*) (Horrocks et al. 2004) qui étend l'ensemble des axiomes OWL en incluant des règles de Horn (*Horn-like rules*). L'interopérabilité avec les ontologies OWL est assurée en référençant les concepts et propriétés OWL à l'intérieur des règles SWRL. SWRL étend les axiomes existants pour OWL Lite et OWL DL afin d'inclure les clauses de Horn. SWRL vient avec toute l'expressivité de OWL DL, mais au prix de la décidabilité et des implémentations pratiques. Il est possible de garder la main sur la décidabilité en limitant la forme des règles admissibles, en imposant une condition de sécurité appropriée (Motik & Sattler 2006).

En effet, dans sa version initiale, SWRL n'est pas décidable et ne permet pas de créer des instances de concepts dans une ontologie. Dès lors, afin de pouvoir continuer à utiliser des raisonneurs, au-dessus d'ontologies OWL comprenant des règles SWRL, l'approche *DL-safe rules* (Motik & Sattler 2006) est une restriction de SWRL qui, lorsqu'implémentée, permet de regagner la décidabilité. Afin de créer un sous-ensemble décidable de SWRL, les règles sont restreintes afin d'agir uniquement sur les instances actuellement présentes dans l'ontologie.

Les règles SWRL ont la forme d'une implication entre un antécédent (ou le corps) et un conséquent (la tête). Une telle règle doit être interprétée comme suit : chaque fois que les conditions spécifiées dans l'antécédent sont valides, les conditions spécifiées dans le conséquent doivent également être conservées. Tant l'antécédent (corps) que le conséquent (tête) sont constitués de zéro ou de plusieurs atomes. Un antécédent vide est considéré comme satisfait par toute interprétation, de sorte que la conséquence doit aussi être satisfaite par toute interprétation; un conséquent vide est considéré comme non satisfait par aucune interprétation, de sorte que l'antécédent non plus ne doit être satisfait par aucune interprétation.

D'après la ***définition 7*** *(Règle ou axiome de règle)*, si l'ensemble des conditions présentes dans l'antécédent sont évaluées à **VRAI**, alors son conséquent est dérivé en tant que nouvelle assertion de la base de connaissances. Une règle avec un antécédent vide sera considérée vraie, de la même manière qu'une règle avec un conséquent vide sera considérée fausse. Lorsque des formules non atomiques sont utilisées dans l'antécédent ou le conséquent, elles sont interprétées en tant que conjonction.

SWRL autorise les atomes ou formules atomiques suivantes:
- **C(e_1)** – VRAI si e1 est une instance de la classe C ou est de type C - atome de classe ;
- **P(e_1;e_2)** – VRAI si la propriété P existe entre les deux instances e1 et e2 ;
- **sameAs(e_1;e_2)** – VRAI si e1 et e2 sont reliées à travers le prédicat owl:sameAs;
- **differentFrom(e_1;e_2)** – VRAI si e1 et e2 représentent deux instances différentes ;
- **buildIn(r;e_1;e_2; :::)** – VRAI si la relation r est VRAI pour les arguments e1, e2,

Les règles SWRL manipulent des variables (précédées par le caractère "?"). Leur syntaxe permet de composer des règles complexes (voir Figure 19). Dans le cadre de mes recherches, j'ai principalement travaillé avec SWRL et ce pour les raisons listées ci-dessous:

- **Standard** - Le langage SWRL est un standard proposé par le consortium W3C. Sa syntaxe est publiée et maintenue par des membres du W3C, et est publiée sur le site du W3C[11]
- **Supporte le raisonnement** - Une fois la restriction *DL-safe* rules appliquée, les algorithmes de raisonnement exécutées au-dessus de bases de connaissances modélisées en OWL (1 ou2) sont capables de prendre en compte les conséquents des règles SWRL.
- **Compréhensible** - Alors que basée sur une syntaxe en notation Extended BNF (très similaires à XML), la spécification SWRL comprend aussi avec une syntaxe adaptée pour les utilisateurs humains, facilitant dès lors l'écriture, la modification et la composition des règles logiques. Dans cette syntaxe,

[11] https://www.w3.org/Submission/SWRL/

une règle a la forme suivante antécédent -> conséquent, antécédent et conséquent étant des conjonctions d'atomes (a_1 ^a_2 ^... ^a_N).

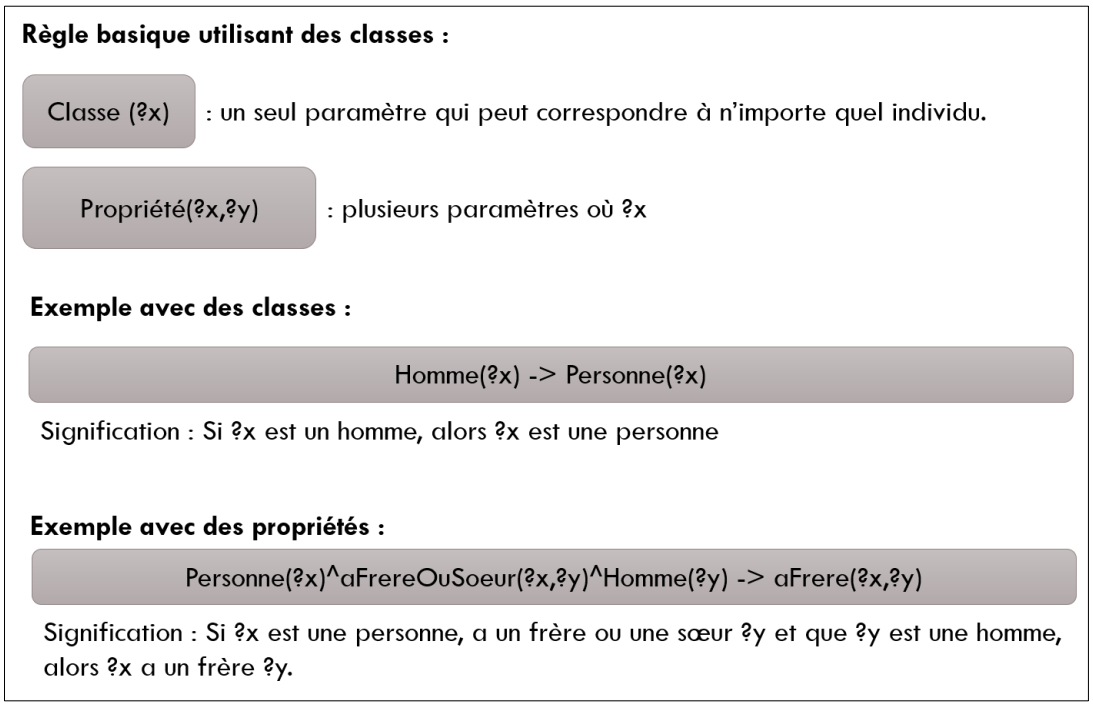


Figure 19. Exemple de syntaxe de règles en SWRL.

3.1.4.3.3 Les raisonneurs potentiels

Lorsqu'une requête SPARQL est adressée à une base de connaissances comportant des clauses de Horn, les langages de programmation logique traditionnels, tels que Prolog, vont employer une méthode de résolution appelée SLD (*Selective Linear resolution for Definite clauses*) (Apt 1997). Cette méthode revient à interpréter la clause la plus haute (à savoir la requête) en tant que négation de la conjonction des différentes sous-requêtes. Une technique de chaînage arrière est alors utilisée (Gallier 2015)(Russell & Norvig 2009).

Lorsqu'on souhaite combiner l'utilisation de clauses de Horn avec des raisonneurs basés sur OWL, 4 principales approches existent:

1. Les règles SWRL peuvent être traduites en logique du premier ordre et dans ce cas les résultats des tâches de raisonnement peuvent être prouvés par le biais de preuves de théorèmes (ou des "*theorem prover*"). Le raisonneur Hoolet implémente cette approche (Bechhofer 2004).
2. Les axiomes OWL-DL peuvent être traduits en des règles, puis passés aux raisonneurs (qu'ils soient basés sur un chaînage avant ou arrière). Cela est implémenté par les raisonneurs Bossam (Jang & Sohn 2004), Kaon2 (Motik & Studer 2005), SWRL-IQ (Elenius 2012), SWRL2COOL (Rigas et al. 2012) et O-Device (Meditskos & Bassiliades 2008).
3. Les règles SWRL peuvent être traduites en des axiomes DL (Parsia et al. 2005). Ces règles sont par la suite traitées par un raisonneur OWL-DL exploitant l'algorithme Tableaux (c'est le cas pour les raisonneurs Pellet (Sirin et al. 2007) et RacerPro (Haarslev et al. 2012)) ou encore l'algorithme Hypertableau (Motik et al. 2009) (c'est le cas pour le raisonneur HermiT (Glimm et al. 2014))
4. Le raisonnement peut aussi être implémenté en utilisant une approche de ré-écriture de requêtes. Dans une telle approche, les règles SWRL sont prises en compte lors de la ré-écriture de requêtes. Le raisonneur utilisé par le magasin de triples Stardog (Inc. 2016).

Enfin il existe des raisonneurs s'appuyant sur des approches hybrides, utilisant des techniques à la fois de chaînage avant et chaînage arrière. C'est le cas des raisonneurs DL2DB (Mei et al. 2006), et Jena (Jena 2016), le dernier ne supportant pas les règles SWRL.

Le tableau suivant liste les principaux raisonneurs, qui, à ma connaissance, supportent à la fois le langage OWL et le langage SWRL et est extrait d'une de nos publications [RIS5].

Raisonneur	**Raisonne à l'exéc. de la requête ?**	**Matérialise le contenu de la ABox ?**	**Expressivité DL**	**Pour grands magasins de triples ?**
KAON2	Oui	Non	$\mathcal{SHIQ}(\mathcal{D})$	Oui
RacerPro	Non	Oui	$\mathcal{SHIQ}$	Oui
Hoolet	Oui	Non	$\mathcal{SHOIN}(\mathcal{D})$	Oui
Pellet 2.0	Non	Oui, tout en supportant le raisonnement incrémental.	$\mathcal{SROIQ}(\mathcal{D})$	Oui
HermiT	Non	Oui. Matérialisation partielle lors de chaque test de cohérence.	$\mathcal{SROIQ}(\mathcal{D})$	Oui
SWRL2COOL & O-Device	Non	Oui	$\mathcal{SHIF}(\mathcal{D})$	Non
SWRL-IQ	Oui	Non	$\mathcal{SHI}$	Oui
DL2DB	Oui	Non	$\mathcal{SHI}$	Oui
Bossam	Non	Oui	$\mathcal{SHI}$	Oui
Stardog	Oui	Non	$\mathcal{SROIQ}(\mathcal{D})$	Oui

Tableau 10. Liste des raisonneurs supportant les règles SWRL. (adapté depuis [RIS5])

3.1.5 Caractéristiques des données sémantiques

Nous venons de définir les principaux composants d'une ontologie comme étant les concepts, les relations, les propriétés et les instances. Ces éléments permettent de constituer un vocabulaire ontologique pour un domaine donné. Une ontologie peut alors être vue comme un ensemble de propositions exprimées utilisant les termes de ce vocabulaire.

A première vue, modéliser une ontologie semble être similaire à la modélisation de logiciels orientés objet ou encore à la conception de diagrammes de type entité-relation pour les bases de données. Il y a cependant deux différences majeures (Grimm et al. 2011) :

> Premièrement, les langages de description d'ontologies fournissent des sémantiques formelles plus riches que celles fournies par les formalismes orientés-objet ou liés aux bases de données. Une ontologie spécifie une "axiomatisation" de connaissances dans un certain domaine, plutôt qu'un modèle de données ou qu'un modèle objet.
> Deuxièmement, les ontologies sont généralement développées à des fins différentes par rapport aux modèles orientés-objet ou par rapport aux diagrammes entités-relations. Alors que les derniers décrivent des composants d'un système d'information, ou alors un schéma pour le stockage de données, une ontologie capture la connaissance d'un certain domaine en tant que telle et permet de raisonner à propos de cette connaissance.

Les technologies sémantiques fournissent des moyens génériques et flexibles favorisant la découverte et l'intégration de données à partir de sources de données réparties sur le Web. Elles présentent notamment les avantages suivants lorsque comparées aux approches traditionnelles de modélisation et d'intégration de données (e.g. API, bases de données relationnelles):

> **Approches orientées données** - Traditionnellement, la modélisation des données démarre avec la définition d'une structure (schéma) qui est ensuite instancié par l'utilisateur avec des données réelles (issues du Web, d'applications, de bases de données, etc.). Avec les approches sémantiques il est toujours possible de fonctionner ainsi, mais il est surtout possible d'inverser ce processus. Il est en effet possible de partir directement des données (indépendamment de leur taille ou niveau

d'hétérogénéité) et soit manuellement, soit automatiquement, les affecter à des classes (soit définies au préalable soit générées à partir des données elles-mêmes si couplage avec des approches machine learning).

- **Multi méta-niveaux** - Le modèle de données RDF est un langage abstrait pouvant être utilisé pour définir des instances de données, des structures de données ainsi que les relations entre elles. Dans d'autres approches, deux langages auraient été nécessaires pour cela. Par exemple, dans l'approche STEP (ISO 10303), on utilise le langage EXPRESS pour les structures de données et le langage SPFF (*STEP Physical File Format*) pour les instances de données. RDF permet de combiner des déclarations à plusieurs niveaux à travers son concept de plus haut niveau, à savoir **`rdf:Resource`**, qui peut être défini à tout méta niveau. En effet, rdf:Resource peut être relié grâce à la relation rdf:type à une instance, une classe, une méta-classe, etc. Cette fonctionnalité abstraite de RDF rend ce modèle générique et lui permet de d'être un candidat idéal pour l'intégration de données hétérogènes.
- **Classes indépendantes des propriétés** - Dans les approches de modélisation traditionnelles (y compris pour les technologies STEP e.g. EXPRESS/SPFF/SDAI), on définit un méta-concept de base auquel sont reliés les autres concepts. Par exemple, en EXPRESS, le méta-concept de base est le concept **`Entity`** qui implémente des attributs et des contraintes. La définition d'un attribut se fait toujours dans le contexte d'un concept **`Entity`**. Or, ce n'est pas le cas avec les approches sémantiques, où les classes et les propriétés d'une ontologie sont indépendantes entre elles. Il est dès lors possible de spécifier des propriétés et contraintes en dehors de toute classe et de les associer à différentes classes.
- **Monde ouvert** (*Open World Assumption*) - L'absence d'information dans une ontologie n'est jamais interprétée comme négative, comme c'est le cas dans les approches "monde clos" (*Closed World Assumption*). Par exemple, si on spécifie qu'Alice est assise à côté de Bob, puis qu'on exécute la requête "à côté de qui Alice est assise?", le raisonneur ne pourra pas retourner une réponse. En effet, Alice peut être assise à côté d'autres personnes, et cette connaissance n'a seulement pas été spécifiée dans l'ontologie. Le fonctionnement en monde ouvert est une conséquence directe de l'idée à la base du web "tout le monde peut tout dire sur tout". Il est toutefois possible de forcer un fonctionnement en monde clos, et dans ce cas il faudrait spécifier de manière explicite qu'Alice n'est pas assise à côté d'autres personnes.
- L'**identification des ressources à base d'URIs** permet de réutiliser ces URIs pour définir des liens entre les données en créant un espace global de données pouvant être parcouru par un agent informatique.
- **Pas de nom unique** (*No Unique naming assumption*) - Ce n'est pas parce que deux concepts ont le même nom qu'un raisonneur déterminera qu'ils sont équivalents. La différence entre concepts doit être exprimée de manière explicite. Si l'on considère deux instances de la classe Personne, "Alice" et "Bob", elles peuvent potentiellement faire référence au même individu. Si ce n'est pas le cas, il faut spécifier dans l'ontologie que l'"Alice" est différente de l'instance "Bob". A titre de rappel, deux concepts sont considérés identiques par un raisonneur si et seulement si leurs identifiants (en d'autres termes les chaînes de caractères composants leurs URIs) sont identiques. OWL fournit cependant des prédicats permettant de spécifier des "égalités" entre classes, propriétés et instances (respectivement **`owl:equivalentClass`**, **`owl:equivalentProperty`** et **`owl:sameAs`**).
- **Normalisation ultime** - Un modèle RDF est un graphe orienté et étiqueté de ressources identifiées par le biais d'URIs. Les ressources sont contenues dans des triplets respectant la forme <Sujet Prédicat Objet>. Un triplet n'est jamais modifié, seulement ajouté ou supprimé. Les triplets peuvent donc être vus en tant qu'atomes de tout modèle RDF. Lorsque l'on fusionne plusieurs modèles RDF, cela revient à placer ensemble les triplets les constituant (tout en vérifiant la consistance logique). Lorsque l'on compare ce fonctionnement aux formes normales en modélisation de bases de données, RDF fournit une forme normale maximale. Les magasins de triplets implémentant RDF dé-normalisent à nouveau (matérialisation des triplets) et ce afin d'augmenter les performances en lecture.
- L'**accès aux données via le protocole HTTP** universel permet l'exploration de toute source de données ayant un emplacement réseau et supportant la négociation de contenu
- **Inférence logique** - Les langages RDFS et OWL étant basés sur un fragment de la logique du premier ordre (FOL), les données ne sont pas seulement assertées (explicitement ajoutées à la ABox) mais elles peuvent être déduites ou inférées grâce à l'utilisation de raisonneurs implémentant les sous-ensembles nécessaires de la logique du premier ordre.

Les ontologies définies en utilisant les langages standards du Web sémantique sont conçues afin d'être spécifiés, étendues et maintenues :

> de manière distribuée
> de manière incrémentale ou progressive

Il s'agit en effet de conséquences directes de l'hypothèse du monde ouvert. Partons de l'exemple d'une ontologie spécifiée avec un langage basé sur la logique de description. Au début, elle est vide, donc à priori tout est possible. C'est lors du processus (itératif) de modélisation, que des contraintes sont ajoutées ce qui rend l'ontologie plus restrictive. Lorsque l'on travaille en monde ouvert, il s'agit de spécifier ce qui n'est pas possible, interdit ou exclu. Cela correspond à comment nous, êtres humains, fonctionnons dans le monde réel: nous avons l'habitude de gérer les informations incomplètes. C'est pour cela que dans le Web sémantique, les choses ne sont pas bien différentes. La modélisation des connaissances est obligatoirement un processus itératif car les informations dont nous disposons sont incomplètes. Le modèles ontologiques définis sont amenés à être étendus ou restreints, selon comment notre connaissance évolue. Et il nous est impossible de prévoir à l'avance comment un tel modèle va évoluer. Le raisonnement en monde ouvert suppose donc des informations incomplètes par défaut (ou au départ) et c'est la raison pour laquelle il est fréquent d'avoir des modèles volontairement sous-spécifiés; ces modèles seront repris et étendus en fonction de l'évolution des connaissances du modélisateur. En effet, un système fonctionnant en monde clos requiert une spécification complète des connaissances. Contrairement aux systèmes fonctionnant en monde ouvert, dans un système clos on spécifie de manière explicite ce qui est possible. Dans un tel système, tout ce qui ne peut pas être déduit (ou prouvé **`VRAI`**) à partir des connaissances spécifiées sera considéré comme faux.

3.2 Profilage utilisateur avec gestion de l'incertitude

3.2.1 Perspective historique

Ce premier chapitre est consacré à la gestion de l'incertitude avec des ontologies. Il traitera de notions sous-jacentes comme la classification et la catégorisation.

La modélisation du contexte et son intégration pour délivrer des approches sensibles au contexte a longtemps été une problématique centrale dans mes recherches, et ce depuis ma thèse de doctorat. Ayant travaillé sur un protocole de découverte de services Web sémantiques sensibles au contexte de l'utilisateur, j'ai dû étudier les approches existantes permettant d'approximer un contexte (d'utilisateur ou d'utilisation). J'avais déjà identifié les approches à base d'ontologies comme pertinentes pour la modélisation du contexte mais aussi pour implémenter une sensibilité à celui-ci de la part d'un composant informatique.

Ce n'est que tout naturellement que j'ai commencé à travailler en 2012 avec Madame Anett HOPPE sur un problème de recherche similaire: comment arriver à profiler des utilisateurs à travers leur navigation sur le Web et comment leur proposer des publicités adaptées à leur profil ? A la suite de la signature d'un contrat de thèse CIFRE N°2012-0926 avec la société ezakus (Bordeaux, France), et sous la direction du Prof. Christophe NICOLLE, j'ai co-encadré les travaux de Madame HOPPE sur le sujet. Madame HOPPE a soutenu sa thèse en février 2016, et depuis travaille en tant que chercheur à la Bibliothèque Technique Allemande (TIB) à Hanovre en Allemagne. Nous continuons d'échanger et de collaborer.

L'entreprise ayant financé ces recherches, ezakus, est une entreprise du domaine de la publicité en ligne. Elle est spécialisée dans la fourniture de services aux acteurs de la publicité en ligne, services allant du stockage et l'organisation de données à du profilage spécifique et de l'optimisation de campagnes publicitaires.

Le développement des technologies pour la publicité en ligne a été largement dominé par des acteurs issus du domaine du *machine learning* (apprentissage automatique), et pas seulement en France. En effet, les données publicitaires présentent deux caractéristiques qui les rendent particulièrement intéressantes pour les chercheurs en machine learning: elles sont très riches, et elles permettent d'implémenter des prédictions et des analyses prédictives. C'est la principale raison pour laquelle la majorité des entreprises offrant des services d'analyse de données dans le domaine de la publicité en ligne utilisent des techniques d'apprentissage automatique. Parmi celles-ci, nous pouvons citer les réseaux neuronaux (Murray & Durrell 2000), les machines à vecteurs de support (eXelate 2012) ou encore les arbres de régression (Vogel et al. 2007). En effet, des chercheurs de renom affirment que "des modèles simples et beaucoup de données surpassent les modèles plus élaborés basés sur moins de données" (Halevy et al. 2009). Les auteurs défendent l'idée selon laquelle lorsque les ensemble d'apprentissage atteignent une taille suffisamment grande (millions voire milliards d'instance de données), les modèles basiques (comprenant des hypothèses simples) permettent d'obtenir de meilleurs résultats que des modèles plus complexes (e.g. des ontologies conçues manuellement). Ceci a permis d'avoir une croyance générale selon laquelle les méthodes d'apprentissage automatique représentent l'unique solution pour des contextes d'applications impliquant beaucoup de données. Des techniques plus sophistiquées comme par exemple l'analyse sémantique ou la modélisation sémantique de domaines de connaissances sont souvent discréditées, souvent de par leur complexité et leur coût (installation, calcul, maintenance).

Toutefois, l'apprentissage automatique a ses limites. La plupart des techniques sont basées sur des représentations des données en entrée sous forme de vecteurs. Des ressources complexes sont souvent

réduites à un vecteur de mots-clés. Je ne rediscuterai pas ici les limites des représentations à base de mots-clés (Roxin 2009). Il apparaît comme évident que des concepts aussi complexes que des préférences utilisateur ou des documents en langage naturel, sont difficilement "représentables" à travers des mots-clés. Les ambiguïtés introduites par les relations linguistiques entre termes (e.g. synonymie, hyponymie) suffisent à justifier ce point. Mis à part ces problèmes liés à la structure du langage naturel, il existe d'autres caractéristiques des systèmes d'apprentissage automatique qui peuvent provoquer des effets de bord indésirables dans un contexte industriel. En effet, la majorité des systèmes *machine learning* se basent exclusivement sur des ensembles d'apprentissage pour la création de leur modèle de classification interne. C'est ce modèle qui est ensuite utilisé pour classifier les instances de données entrantes. Il est très rare que l'algorithme désigne une adaptation de ce modèle suite à la détection d'une modification dans la distribution des données. Les modifications sur les données ont souvent pour conséquence de ré-entraîner et de réaccorder l'ensemble du système. Dans un tel mode de fonctionnement, on ne vise pas l'intégration de connaissances métier. L'analyste qui peut uniquement influencer les variables utilisées en entrées, pas leur importance ou leur combinaison en vue de la classification finale. Avec l'apprentissage automatique, il est en général impossible de spécifier des connaissances métier expert et ce de manière à les intégrer durant le processus de classification. Certaines approches semi-supervisées, permettant l'insertion de règles de classification restreintes, seraient une exception pour cette règle. Toutefois, nous n'avons pas trouvé de références dans la littérature discutant d'une potentielle intégration dans un système de profilage industriel. De plus, le résultat d'un processus d'entraînement, e.g. le modèle de classification interne du système, s'apparente généralement à une boîte noire. Le système lit les données fournies en entrée, puis fournit un résultat de classification. Le système ne fournit aucune justification, n'expose aucunement le processus de raisonnement utilisé pour arriver au résultat fourni en sorti (la classification proprement dite). Il n'est donc pas possible de comprendre pourquoi un certain résultat a été produit par le système. Or, de telles informations pourraient fournir des aperçus intéressants sur a) l'évolution de la distribution des données, b) l'importance des variables en entrée et c) les éventuelles erreurs dans la construction du modèle de classification.

Etant données ces limites, nous avons souhaité, à travers les recherches poursuives dans le cadre de cette thèse, explorer les avantages offerts par les technologies sémantiques en comparaison aux approches basées sur de l'apprentissage automatique, ainsi que leur potentielle intégration dans un système industriel (Bürger & Simperl 2008).

3.2.2 Description du domaine et des problèmes étudiés

Le but des recherches menées au sein de cette thèse fut de concevoir une nouvelle approche pour un système de profilage utilisateur. Le but des systèmes de profilage dépend, dans une large mesure, de l'application au sein de laquelle il s'intègre ou qui exploite ses résultats. Dans les travaux présentés ici, notre domaine d'application est celui de la publicité en ligne.

Or, avec l'extension du marché publicitaire aux médias numériques, des changements considérables sont apparus par rapport à l'écosystème de la publicité sur les médias traditionnels. En effet, les emplacements publicitaires ne sont pas échangés, au travers de négociations directes, entre l'annonceur et l'éditeur. Au lieu de cela, les emplacements publicitaires et leur public sont analysés automatiquement et vendus aux enchères. Ceci se déroule au travers de plateformes automatisées de vente et d'achat d'espaces publicitaires, en temps réel. L'éditeur offre son espace publicitaire dans un environnement d'enchères. Les annonceurs placent ensuite des enchères sur les espaces proposés en temps réel. Un emplacement publicitaire proposé sur une telle plateforme peut être caractérisé. En exploitant des traces de navigation d'utilisateurs, les interactions passées d'utilisateurs individuels sont connues et peuvent être utilisées pour faire des déductions sur leurs intérêts généraux et leur comportement en tant que clients. En conséquence, les annonceurs placent rarement des enchères sur des espaces publicitaires faiblement caractérisés. Ils préfèrent acheter des emplacements vis à vis desquels ils ont plus de certitudes d'être "livrés" à des clients spécifiquement profilés. Plusieurs entités gravitent dans cet écosystème, à savoir:

> Les plateformes DSP (*Demand Side Platform*) qui permettent aux annonceurs de placer leurs enchères
> Les plateformes SSP (*Supply-Side Platform*) qui permettent aux éditeurs de proposer aux enchères des emplacements libres, au fur et à mesure qu'ils se libèrent (en temps réel)

Dans cet écosystème d'échange de publicités, des connaissances détaillées sur l'utilisateur représentent une ressource importante. Il existe des entreprises spécialisées, qui récupèrent des informations utilisateur

provenant de diverses sources e.g. des sites sur le Web, à partir de données non numériques, par exemple provenant d'interactions clients réelles ou de données de cartes de crédit. D'un point de vue annonceur, ces données sont considérées des données tierces (*third party data*). Ces données sont par la suite couplées avec d'autres données, de l'annonceur par exemple, au sein de plateformes de gestion de données ou DMP (Data Management Platforms). Leur but est de permettre aux annonceurs de gérer leurs données efficacement, en les consolidant en combinant différentes sources e.g. niveaux d'audience, programmes marketing, détails sur les campagnes publicitaires, etc.

Ce type d'approche marketing n'est, bien évidemment, pas réalisable avec des outils manuels. L'utilisation combinée de techniques d'apprentissage automatique et de techniques d'analyse de données Big Data a permis l'avènement de l'écosystème décrit ci-dessus. Toutefois, le but ultime de la publicité en ligne serait de pouvoir cibler chaque utilisateur individuellement. Cela suppose que pour chaque utilisateur, il faut pouvoir être capable de déterminer si son profil fait partie groupe d'audience déterminé au préalable par l'annonceur, pour une publicité donnée. En effet, dans le contexte du projet MindMinings, pour chaque campagne publicitaire, des segments marketing sont identifiés au préalable e.g. la meilleure audience pour ce type de shampooing est constituée par les mamans sportives. Une fois le segment marketing défini (e.g. mamans sportives), il s'agit d'identifier l'ensemble des utilisateurs individuels ayant la plus grande probabilité d'appartenir à ce segment. Les utilisateurs ainsi sélectionnés seront inclus dans la campagne publicitaire, et verront donc la publicité du shampooing en question. Ce ne sera pas le cas pour les utilisateurs dont le profil correspondait dans une moindre mesure au segment marketing.

Souhaitant proposer un système de profilage correspondant en tout point au processus ci-dessous, les exigences du domaine ont fortement impacté nos recherches. Le système de profilage MindMinings a été conçu afin d'intégrer ces exigences en termes d'efficacité de traitement et d'exploitation des entrées du système, et ce afin de délivrer des conclusions pertinentes d'un point de vue marketing. Notre objectif fut d'exploiter de nouvelles techniques pour l'analyse de contenus Web et pour la modélisation d'informations, afin d'augmenter le niveau de connaissance des utilisateurs individuels.

Les travaux menés dans ce contexte se placent donc à la frontière de plusieurs domaines, notamment le ***Web sémantique,*** l'***extraction de connaissances*** et les ***données liées ouvertes*** (*Linked Open Data*). Un des défis associés à ces recherches fut leur application dans un milieu industriel. En effet, alors que les technologies du Web sémantique étaient déjà utilisées dans de nombreux domaines comme par exemple le profilage utilisateur au sens large, elles étaient souvent intégrées en tant que bases de connaissances plutôt "passives" qui étaient exploitées avec des techniques d'apprentissage automatique. A notre connaissance, il n'existait en 2012, à l'échelle industrielle, aucun système de profilage d'utilisateurs, exploitant uniquement des traces de navigation (extraites à partir de cookies).

Un des objectifs de cette thèse fut donc de déterminer la "véracité" de l'idée communément acceptée selon laquelle les approches à base de technologies sémantiques sont, non seulement, peu efficaces, mais en plus présentent un coût important. Nous nous sommes donc intéressés aux ***adaptations possibles à appliquer aux systèmes à base d'ontologies, afin d'améliorer leur utilisation dans un environnement impliquant une utilisation intensive de données***. Des adaptations ont été explorées d'une part au niveau de la conception du système, à travers l'utilisation de modules, et, d'autre part, au niveau du contenu du profil utilisateur, à travers une spécialisation du profil construit pour le domaine de la publicité en ligne. Les résultats ont été évalués au travers de deux axes de performance, à savoir d'un point de vue qualitatif (pertinence des résultats) et d'un point de vue quantitatif (en termes de vitesse de traitement et passage à l'échelle).

Un deuxième objectif poursuivi au travers des recherches décrites dans ce chapitre concerne l'intégration de l'ontologie dans le système. L'état de l'art nous a fourni une liste extensive d'approches de profilage utilisateur à base d'ontologies (voir section suivante). Alors que le degré d'expressivité offert par les ontologies est souvent exploité pour obtenir des représentations de profils utilisateurs plus détaillées (que des représentations traditionnelles), les ontologies en tant que tel sont souvent intégrées comme élément de stockage des informations, avec un rôle plutôt passif dans l'intégralité du système de profilage. Au travers des recherches menées dans le cadre du projet MindMinings, nous avons souhaité ***exploiter au maximum les possibilités de raisonnement offertes*** et d'***utiliser des règles logiques afin d'adapter dynamiquement le contenu du modèle de connaissance***. L'idée a été d'explorer le ***concept de degré de connaissance***. En effet, alors que jusqu'ici, en informatique, les ontologies sont utilisées en respectant le paradigme du monde clos (les connaissances contenues représentent ce qui existe), nous avons investigué la possibilité d'exprimer ce qui pourrait être. Notre idée sous-jacente fut d'associer une valeur de confiance à certaines relations dans

le modèle. Dans un contexte comme celui de la publicité en ligne, où la connaissance d'un utilisateur est construite à partir d'éléments implicites, il nous a apparu important de pouvoir avoir une probabilité associée à certains énoncés. Ainsi, le système pourra calculer la probabilité que présente un profil utilisateur donné d'appartenir à un segment marketing donné.

Enfin, soucieux de modéliser un système de profilage qui soit le plus fidèle à la réalité industrielle, nous avons souhaité intégrer l'ontologie non plus comme simple stockage de connaissances, mais bien en tant que composant dynamique pouvant s'adapter à la grande diversité de segments marketing. L'approche conçue permet de spécifier des segments marketing à travers un ensemble de règles logiques. Ces règles peuvent être construites sans requérir des compétences spécifiques, au moyen d'une interface intuitive développée spécifiquement pour les besoins de ce projet.

Les travaux de recherche présentés dans ce chapitre ont été menés en réponse aux questions de recherche suivantes:

> Déterminer la faisabilité d'un processus de profilage des ressources, intégré à échelle industrielle, basé sur les ontologies
> Est-il possible de modéliser l'ensemble des connaissances nécessaires avec des langages de description d'ontologies standard ?
> Est-il possible d'utiliser un modèle ontologique pour le processus de profilage lui-même et pouvoir utiliser un raisonneur pour construire le profil d'une ressource de manière (semi-)automatique ?
> Est-il possible de concevoir un système avec une maintenance aisée, facile à mettre à jour, permettant d'adapter la procédure de profilage selon les besoins ?
> Évaluer la performance du processus de profilage en termes de qualité (complétude des résultats retournés), de vitesse et passage à l'échelle
> Évaluer l'apport des ontologies dans le profilage utilisateur (notamment impact du raisonnement sur l'efficacité)
> Déterminer l'applicabilité des ontologies à un tel écosystème
> Est-ce qu'une approche à base d'ontologies permet de répondre à toutes les contraintes du domaine ?
> Quel est le niveau de performance atteint par notre approche en comparaison avec l'approche machine learning implémentée par l'entreprise ?

Notre problématique initiale a été de déterminer si, avec des méthodes modernes d'extraction de connaissances, il est possible d'***améliorer la compréhension que nous avons d'une page Web représentée uniquement à travers son URL***. Une fois de telles connaissances extraites à partir de la page Web elle-même, nous pensions utiliser les technologies du Web sémantique afin d'intégrer ces connaissances dans une ontologie plus complexe. C'est ce noyau ontologique qui nous permettra d'implémenter le profilage d'un utilisateur, mais aussi de pondérer un profil utilisateur par rapport à un segment marketing.

Toutefois les recherches ont été fortement impactées par le domaine spécifique de l'entreprise, spécialisée dans la publicité en ligne, et les demandes spécifiques en lien, notamment:

> Pouvoir "classer" par rapport à une taxonomie de catégories les pages Web consultées par un utilisateur donné
> Pouvoir composer à la volée des segments marketing en utilisant des concepts et propriétés de l'ontologie - l'entreprise était en charge de campagnes de publicité, et ces campagnes nécessitaient d'identifier des segments marketing, à savoir des groupes d'utilisateurs ayant des caractéristiques communes (e.g. le segment marketing des "mamans sportives")
> Pouvoir associer une probabilité qu'un profil utilisateur appartienne à un segment marketing donné - les publicités étant associées à des segments marketing regroupant plusieurs utilisateurs, l'entreprise souhaitait pouvoir classer les utilisateurs d'un segment marketing par rapport à la probabilité qu'ils y appartiennent bien (si l'on reprend le segment des "mamans sportives", il est évident que si un utilisateur se retrouve associé à ce segment sans correspondre en tout point au profil souhaité, il ne sera pas impacté par la publicité)
> Comparer, en termes d'efficacité et de pertinence, deux approches différentes: d'une part celle basée sur les technologies du Web sémantique (poursuivie à travers les travaux de thèse de Madame Hoppe), d'autre part une approche basée sur du machine learning (implémentée dans les locaux de l'entreprise, par ses équipes).

Dans ce qui suit, je présente un résumé de l'état de l'art du domaine du profilage utilisateur dans un environnement Web, en exploitant uniquement ses traces de navigation. Les approches de profilage utilisateur dans un environnement Web, mais utilisant des données explicites, n'ont été que rapidement parcourues et ne seront pas discutées dans les paragraphes suivants.

3.2.3 Rappel de l'état de l'art spécifique à la thématique considérée

Les recherches présentées ici traitent du profilage d'utilisateurs à partir de l'analyse des pages Web consultées. Il convient dès lors de présenter un bref état de l'art du profilage d'utilisateurs sur le Web, plus particulièrement des approches existantes permettant une "personnalisation" dans un contexte Web. Le lecteur souhaitant plus d'informations sur les approches de profilage en général pourra consulter les références suivantes:

> le profilage pour obtenir des retours d'informations personnalisées (Ghorab et al. 2013)
> état de l'art des approches de profilage pour les systèmes de recommandations (Chu & Park 2009)
> accès personnalisé à l'information (Gauch et al. 2007)
> recherche personnalisée d'informations (Micarelli et al. 2007)

Le "Profilage utilisateur" est souvent considéré compris de manière intuitive, c'est la raison pour laquelle peu d'articles en donnent une définition claire. Quelques exemples se retrouvent dans (Mizzaro & Tasso 2002) où la "personnalisation du web" est définie en tant que "processus de sélection, préparation et fourniture de contenus Web à un utilisateur donné, en prenant en compte ses besoins spécifiques et préférences[12] (Hoppe 2016).

Si on cherche une définition plus claire pour un "profil utilisateur", nous sommes confrontés au même problème - très peu d'auteurs ont pris la peine de donner une définition explicite de ce concept dans leurs publications. Nous pouvons citer (Lops et al. 2011) qui définit un "profil utilisateur" comme "une représentation structurée des intérêts de l'utilisateur, adoptée afin de recommander de nouveaux éléments intéressants" (on remarque l'orientation vers les systèmes de recommandation). Devant le peu de définitions trouvées par rapport à ces concepts, nous avons choisi d'adapter cette dernière définition au contexte de nos travaux. Pour nous, un profil utilisateur est une "représentation structurée d'informations individuelles concernant l'utilisateur en vue de lui offrir des services personnalisés[13]" (Hoppe 2016).

Lorsqu'on s'intéresse aux applications de profilage, les buts poursuivis sont divers et variés, allant au-delà des applications Web. Etant donnée cette variété, nous n'avons pas établi un état de l'art exhaustif de l'ensemble des systèmes de profilage existants. Nous nous sommes particulièrement intéressés aux systèmes de profilage qui, d'une part, basent leur analyses sur des données extraites du Web (en observant le comportement de l'utilisateur et non pas à travers des questionnaires), et, d'autre part, utilisent des modèles explicites pour représenter le profil de l'utilisateur.

Afin de déterminer comment orienter nos recherches, le processus de profilage a été divisé en plusieurs étapes (voir Figure 20), et pour chacune de ces étapes nous avons étudié d'une part comment l'entreprise traitait le problème, et d'autre part quelles contraintes cela soulevait par rapport à l'approche globale. Une fois cette étude réalisée, nous avons évaluées les approches présentes dans la littérature traitant d'un contexte d'application proche ou similaire, à savoir la construction d'un profil utilisateur à partir de traces de navigation implicites.

[12] « process of selecting, preparing and delivering web contents for a given user, by taking into account his specific needs and preference »
[13] « structured representation of individual user information with the goal of offering personalised services. »

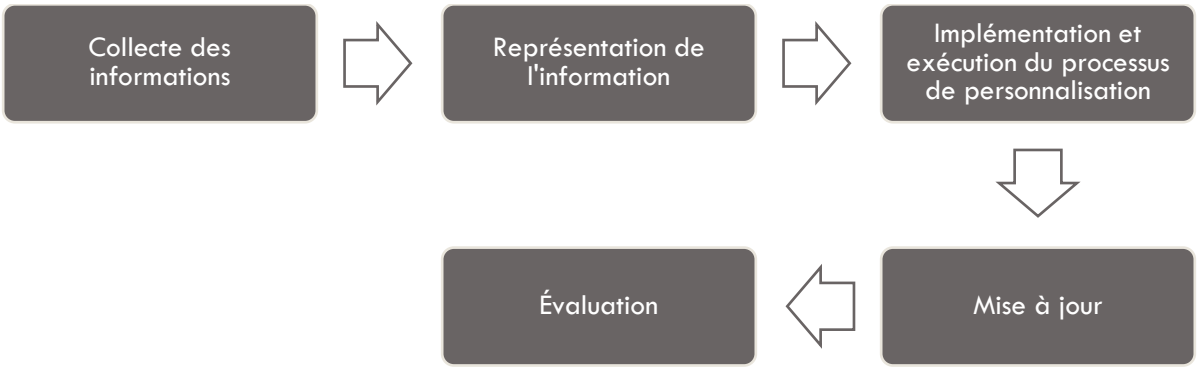


Figure 20. Etapes de la construction d'un profil utilisateur, adapté d'après (Ghorab et al. 2013).

Inspirée de la recherche d'informations, cette division en étapes a l'avantage d'inclure des mécanismes pour exploiter et évaluer le profil utilisateur construit auparavant. Malheureusement il ne prend pas en compte la mise à jour du profil.

Dès lors nous avons intégré ces différentes étapes afin de constituer une vue schématique du processus de profilage tel qu'implémenté avec notre approche. Nous avons ensuite réalisé un état de l'art des approches employées pour chaque étape; j'en présente un résumé dans ce qui suit.

3.2.3.1 Collecte d'informations

Au niveau de la phase de "collecte d'informations", nous nous sommes concentrés sur les approches basées sur des informations implicites. Nous n'avons pas pris en compte les approches recueillant des informations explicites e.g. à travers des questionnaires, formulaires ou sondages. Lorsque l'on considère la collecte d'informations implicites, l'effort analytique à fournir lors de l'interprétation des données est bien plus important que dans le cas d'une collecte d'informations explicites. De plus, les méthodes de collecte d'informations implicites remettent en question la vie privée de l'utilisateur. En effet, dans (Toch et al. 2012), les auteurs remettent en cause l'hypothèse selon laquelle les techniques de collecte côté client offrent un plus grand respect de la vie privée de l'utilisateur. Cependant, pour garantir l'acceptation du système par les utilisateurs, le degré de confidentialité perçu peut être plus important que le degré de confidentialité réellement atteint. Concernant le type d'informations ainsi que les techniques de remontée utilisées par les différentes approches, une étude détaillée est disponible dans (Kelly & Teevan 2003).

Concernant le type d'actions utilisateur pouvant être identifiées, elles sont diverses et concernent : les marque-pages (Parent et al. 2001), (Sieg et al. 2004), le temps passé sur une page (Morita & Shinoda 1994), (Claypool et al. 2001), le défilement (Sakagami & Kamba 1997), (Claypool et al. 2001), la copie de contenu (Morita & Shinoda 1994), l'annotation (Golovchinsky et al. 1999), ou encore le nombre de clics réalisés sur la page (Claypool et al. 2001). Les approches utilisant vraiment des informations de navigation sont très peu nombreuses et visent avant tout l'amélioration de l'expérience utilisateur dans des applications déconnectées du contexte Web (Chirita et al. 2007), (Challam et al. 2007).

Enfin, une étude a été menée afin d'identifier les principales approches permettant de pondérer les informations collectées, lors de leur intégration dans le modèle. Les biais sont les hypothèses centrales qui sous-tendent chaque modèle de données et leur interprétation. Par exemple, toutes les approches présentées dans notre état de l'art font la même hypothèse: toutes sous-entendent le fait que la navigation Web des utilisateurs éclaire sur leur personnalité et leurs comportements futurs. Certaines publications évaluent cette hypothèse en confrontant les participants à l'étude avec, à la fois, un système personnalisé et une version non personnalisée du même système. Leurs résultats confirment notre hypothèse de départ. Dans (Teevan et al. 2005), les auteurs comparent différentes paramétrisations d'un système profilé par rapport à une alternative non profilée du même système. Toutes les variations utilisant la personnalisation sont plus performantes que l'approche non profilée, même si ce n'est que légèrement. Peu de temps après, (Agichtein et al. 2006) comparent un classement de résultats de recherche, uniquement basé sur le contenu, avec ses alternatives personnalisées, et aboutissent à des résultats similaires, tout comme (Mylonas et al. 2008).

Plus les caractéristiques à intégrer dans l'application de profilage sont nombreuses, plus leur intégration dans le profil d'utilisateur devient complexe. Au lieu d'associer à toutes ces caractéristiques la même importance pour le profil de l'utilisateur, de nombreuses approches utilisent des pondérations pour équilibrer le profil ainsi construit. Un exemple souvent rencontré consiste à observer le temps qu'un utilisateur passe sur une page Web donnée et de l'utiliser comme facteur de pondération. L'hypothèse sous-jacente est de supposer que plus l'utilisateur passe du temps sur une page Web, plus elle présente un intérêt fort pour l'utilisateur. Cette approche a été utilisée par plusieurs auteurs: (Sakagami & Kamba 1997), (Parent et al. 2001), (Gauch et al. 2007). Elle présente toutefois plusieurs limitations:

> Le temps nécessaire à la lecture d'un texte peut dépendre de sa longueur et de sa complexité - ce facteur a été étudié par (Morita & Shinoda 1994)
> Il se peut que des fenêtres du navigateur demeurent actives (ou ouvertes), sans que l'utilisateur ne les regarde pas
> La vitesse de lecture "normale" de l'utilisateur, ainsi que ses éventuelles "inhibitions de lecture", ne sont presque jamais prises en compte

Lors de l'intégration de plusieurs sources d'information, il peut être utile de pondérer leurs entrées en fonction de leur fiabilité ou de leur importance. Par exemple, dans (Sakagami & Kamba 1997), les auteurs exploitent une multitude de comportements utilisateur tels la visualisation, le défilement ou encore l'agrandissement de pages Web. L'intégration des différents comportements observés repose sur un système de points: des points "de base" pour la visualisation, des points supplémentaires lorsque le comportement de l'utilisateur dénote un intérêt particulier pour la ressource en question. D'autres travaux exploitent des approches similaires: (Lieberman 1995), (Parent et al. 2001), (Gauch et al. 2007), (Sieg et al. 2004), (Castellano et al. 2007), (Chu & Park 2009).

Les pondérations associées aux différentes sources d'information considérées par une application de profilage permettent d'exprimer un certain biais. Souvent, ces pondérations reposent sur des hypothèses intuitives faites par les chercheurs. Cependant, il arrive bien moins souvent que ces approches soient examinées et validées d'un point de vue expérimental. C'est la raison pour laquelle nous avons souhaité une évaluation à la fois qualitative et quantitative de notre approche (voir 0).

3.2.3.2 Représentation d'informations

Concernant les différentes "représentations" possibles de l'information, nous souhaitions à l'origine limiter notre état de l'art pour nous focaliser sur les approches reposant sur des modèles formels pour représenter le profil de l'utilisateur. Toutefois un état de l'art complet a été réalisé, afin de pouvoir comparer les différentes représentations possibles. En prenant en compte 5 types de représentations (e.g. à base de mots-clés, à base de paires clé-valeurs, à base de concepts, à base de réseaux sémantiques et à base d'ontologies), nous avons défini plusieurs critères d'évaluation, pouvant être regroupés selon deux principaux axes d'étude:

> Capacité à capturer et à correctement communiquer le contenu
    > Contenus sémantiques - évalue le sens associé aux éléments de base du profil utilisateur, capacité à exprimer des sémantiques complexes
    > Structure formelle - évalue s'il est possible de définir des relations entre les éléments de base du profil utilisateur et aussi s'il est possible de les spécifier davantage (relations typées, inverses, etc.)
    > Interprétation utilisateur - évalue la facilité avec laquelle l'utilisateur peut comprendre les éléments de son profil
    > Interprétation machine - évalue la facilité avec laquelle l'ordinateur peut interpréter les informations du profil utilisateur
> Complexité à construire et à exploiter le profil
    > Effort de construction
    > Efficacité de manipulation

Cela nous a permis de construire le tableau suivant (voir Tableau 11) identifiant les approches à base d'ontologies comme les plus pertinentes par rapport aux contraintes de notre domaine d'application.

Critère	Mots-clés	Paires clés-valeurs	Concepts	Réseaux sémantiques	Ontologies
Contenu sémantique	-	+	++	-	+++
Structure formelle	---	-	--	+	+++
Interprétation utilisateur	-	-	+	-	+++
Interprétation machine	---	-	+	-	+++
Effort de construction	++	-	-	+	--
Efficacité de manipulation	++	+	-	+	-

Tableau 11.Comparaison des approches de représentation de l'information. (Hoppe 2016)

Si on considère le degré de complexité des sémantiques pouvant être modélisées avec chaque approche, ces sont les ontologies et les approches à base de concepts qui arrivent en tête. Elles supportent la désambiguïsation (utilisation d'URIs pour l'identification des concepts) ainsi que la spécification explicite de l'ensemble des éléments d'un profil. Les ontologies se différencient des approches à base de concepts à travers l'identification par le biais d'URIs non seulement pour les concepts, mais aussi pour les relations les liant. Enfin, les représentations à base de mots-clés ainsi que les réseaux sémantiques utilisent des chaînes de caractères pour encoder le sens d'un concept, elles ne permettraient pas une interprétation correcte par la machine du sens associé.

Lorsque l'on considère le degré de formalisme de la structure de données sous-jacente à chaque représentation, ce sont les approches à base de réseaux sémantiques et celles à base d'ontologies qui correspondent au mieux aux besoins du projet. Les ontologies se différencient en cela qu'elles permettent d'associer des étiquettes sémantiques (annotations) aux relations définies entre concepts.

Si l'on s'intéresse à l'interprétation des contenus ainsi représentés, les ontologies sont à nouveau le type de représentation qui permet la meilleure interprétation à la fois par les machines et par les humains.

3.2.3.3 Exploitation d'un profil utilisateur

L'étape suivante dans le processus de profilage consiste en l'exploitation proprement dite du profil ainsi constitué. Selon (Pasi 2014), un profilage réussi implique deux phases: premièrement la définition du modèle utilisateur, puis, deuxièmement, le processus de peuplement d'un tel modèle avec des données. Après avoir étudié les différentes modélisations possibles pour un profil utilisateur, nous avons étudié l'ensemble des algorithmes existants permettant d'exploiter un profil utilisateur dans un contexte applicatif e.g. adapter une liste de résultats, filtrer un flux d'informations, etc.

Les algorithmes étudiés se divisent en deux principales catégories: les algorithmes supervisés et les algorithmes non-supervisés.

Au niveau des ***algorithmes supervisés***, les techniques sont nombreuses et se basent principalement sur un ensemble annoté de connaissances et un processus d'apprentissage. A partir de connaissances existantes concernant l'utilisateur (e.g. son profil), de tels algorithmes visent à prédire la pertinence de ressources, nouvelles ou jusque-là non découvertes. Nous avons étudié les approches suivantes:

La classification naïve bayésienne[14] est en effet un des choix les plus populaires en termes de profilage utilisateur. Ce type d'approche est largement utilisé pour prédire si des données nouvelles ou entrantes sont pertinentes par rapport à un ensemble d'intérêts connus de l'utilisateur. En s'appuyant sur les instances d'apprentissage connues, chacune des caractéristiques en entrée est liée à un résultat en sortie, avec une probabilité donnée. Dans un contexte de profilage, une telle caractéristique pourrait être l'apparition d'un mot clé donné dans une ressource Web consultée par l'utilisateur. La probabilité totale du degré de pertinence de la ressource Web pour l'utilisateur considéré est calculée en tant qu'agrégation des

[14] https://fr.wikipedia.org/wiki/Classification_naïve_bayésienne

probabilités associées aux différents termes considérés. En effet, les approches bayésiennes considèrent que les caractéristiques fournies en entrée sont toutes indépendantes entre elles. Pour des ressources textuelles, cela revient à considérer que les termes qui apparaissent à l'intérieur d'un texte sont mutuellement indépendants. Or ceci n'est pas vrai pour les textes écrits en langage naturel, étant donné que les termes d'une phrase sont connectés entre eux à travers une structure syntaxique et un sens sémantique. Toutefois, les approches utilisant la classification bayésienne naïve ont démontré des résultats satisfaisants - (Mladenic 1996) a comparé la performance de la classification bayésienne naïve à celle de l'algorithme des K plus proches voisins (*K-Nearest-Nieghbor*), et ce dernier n'a que de très peu dépassé l'approche bayésienne; d'autres études comparatives démontrent que même si la classification bayésienne ne fait pas partie des méthode les plus performantes (Pazzani et al. 1996), elle n'est pas non plus aux dernières places du classement. De plus, ces approches peuvent être appliquées dans un contexte graphe. (Dudev et al. 2008) ont utilisé la classification bayésienne naïve dans un contexte de personnalisation des requêtes, plus particulièrement pour l'ordonnancement de résultats de requêtes SPARQL.

Les réseaux bayésiens[15] ne considèrent pas chaque caractéristique entrée comme indépendante, mais intègrent ces caractéristiques au sein d'un réseau d'événements liés. A l'intérieur de ce réseau, chaque nœud est représenté par un événement possible (e.g. l'occurrence d'un concept clé dans un document de l'utilisateur) alors que les arcs représentent les liens entre ces événements et ont pour étiquette la probabilité que ces deux événements se produisent au même moment. Les réseaux bayésiens sont fortement utilisés dans le mining, l'analyse de risque ou encore la détection de spams. Dans un contexte profilage utilisateur, ces approches ne sont pas tout à fait adaptées- il faudrait connaître le scénario au sein duquel l'utilisateur évolue pour pouvoir obtenir des résultats intéressants. Or, il est difficilement envisageable d'avoir un modèle des événements suffisamment générique afin de pouvoir correspondre à la grande diversité de situations pouvant arriver lorsque l'on considère la navigation sur le Web. (Eyharabide & Amandi 2012) ont présenté une application des réseaux bayésiens dans un contexte navigation Web. Ils ont utilisé ce type d'approche pour découvrir la réapparition d'attributs sémantiques contextuels dans la navigation de l'utilisateur. Toutefois dans ce cas, le contexte de l'utilisateur était limité à un ensemble de relations causales entre certains terminaux et des préférences utilisateur.

Les approches à base des plus proches voisins représentent des approches intuitives, qui emploient une comparaison avec des instances connues afin de déterminer des attributs inconnus de points données entrants. Historiquement basées sur l'algorithme de Rochhio, ces méthodes comparent les instances entrantes à des points de données existants dans la base, et les associent au plus similaire d'entre eux. La méthode des k plus proches voisins peut être vue comme une extension de l'algorithme de Rocchio, car au lieu d'utiliser une seule instance voisine pour déterminer l'appartenance à une classe, cette méthode utilise un ensemble d'instances voisines. La taille de l'échantillon à considérer est déterminée par les paramètres d'entrée de l'algorithme. Selon le contexte d'application, l'appartenance de classe déduite en sortie peut être principalement impactée par un vote majoritaire (dans le cas des classes quantitatives) ou par une agrégation des valeurs des voisins (dans le cas de valeurs numériques). L'algorithme de Rocchio et l'approche de k plus proches voisins sont toutes deux considérées dans l'étude de (Pazzani et al. 1996), la dernière d'entre elles ayant les pires performances.

Les approches à machines à vecteurs support sont des méthodes d'apprentissage utilisant des fonctions noyau pour classifier des données non linéaires. La fonction noyau transfère les points de données dans un espace vectoriel de plus grande dimension, tout en préservant leurs distances relatives. Ceci permet d'appliquer des méthodes de classification plus simples sur les points de données ainsi transformés. Toutefois le choix de la fonction noyau n'est pas trivial et doit être adapté au domaine du problème considéré ainsi qu'à son espace de données. De plus, suite à la transformation inverse, les résultats peuvent être moins "clairs" pour l'interprétation. Dans la littérature on trouve des exemples d'utilisation de cette méthode pour: a) collecter des informations implicites sur l'utilisateur et prédire la probabilité d'achat d'un produit donné (Pandey et al. 2011); b) reconnaître des motifs et facteurs pertinents dans le comportement de l'utilisateur pour l'adaptation d'environnements intelligents (Vildjiounaite et al. 2007); c) recommander des contenus selon les expressions de l'utilisateur (Arapakis et al. 2009).

[15] https://fr.wikipedia.org/wiki/Réseau_bayésien

Les approches à base de règles experts reposent sur des règles construites manuellement par des experts du domaine pour interpréter les données en entrée. La construction d'une telle base de règle est une tâche fastidieuse, c'est pour cela qu'une réponse technologique à cet inconvénient consiste à extraire automatiquement de telles règles. Cela a donné naissance aux approches dites d'extraction de règles (ou rule mining) qui permettent de découvrir des modèles récurrents à partir d'un ensemble de données d'entraînement connues et de formuler les règles correspondantes. Une des principales applications de ce type d'approche concerne l'analyse du panier d'achat sur les sites et applications e-Commerce (Eyharabide & Amandi 2012). (Géry & Haddad 2003) discutent l'application des algorithmes d'extraction de règles à l'étude de traces de navigation Web. Les auteurs y présentent trois approches différentes, chacune étant évaluée selon le "degré de complétude" des extractions de règles, de séquences fréquentes et de séquences à fréquence généralisée.

Les algorithmes évolutionnaires (ou *evolutionary algorithms*) s'inspirent de l'évolution des espèces et des mécanismes naturels de reproduction, mutation, combinaison et sélection. Apparus au tout début des années '60, ces algorithmes produisent un ensemble de solutions, appelé la population. Chaque individu de la population est un candidat à la solution. Pour identifier lequel d'entre eux répond le mieux au problème concerné, chaque individu de la population est évalué par rapport à une certaine fonction de qualité (ou *fitness function*), qui mesure leur performance par rapport au problème. Les individus de la population qui réussissent le test précédent entrent dans la phase de reproduction, afin de générer la génération suivante. Par rapport au domaine qui nous concerne ici, on peut citer deux approches intéressantes reposant sur l'utilisation d'algorithmes évolutionnaires. (Moukas 1997) présente le système Amalthea qui utilise des agents informatiques et des algorithmes génétiques pour assister un utilisateur dans la recherche d'informations. A chaque étape de reproduction, les agents formulent une prédiction concernant les actions futures de l'utilisateur. Ceux qui se rapprochent le plus des actions réelles de l'utilisateur sont sélectionnés pour la phase suivante. (Bouchachia et al. 2014) proposent un algorithme génétique étendu qui, à partir d'une population de vecteurs de centres d'intérêts, permet un calcul incrémental afin que les changements d'intérêt de l'utilisateur soient reflétés dans la population de vecteurs considérée.

Les approches se basant sur l'analyse de liens (*link analysis*) interprètent les relations entre les utilisateurs et leurs centres d'intérêts respectifs à travers une structure de réseau. Les utilisateurs et les centres d'intérêt sont modélisés en tant que nœuds du réseau, et les relations entre eux sont des arcs annotés. Le modèle de liens apparaissant à travers un tel graphe détermine le degré d'intérêt qu'un utilisateur aura pour un domaine donné. Twitter utilise une approche similaire pour son algorithme "Who to follow" (Qui suivre) présenté dans (Gupta et al. 2013). Sur Twitter, chaque utilisateur spécifie ses intérêts de manière explicite à travers les Tweets auxquels il/elle a souscrit. L'analyse de ces abonnements permet de construire un graphe des utilisateurs et de leurs centres d'intérêts qui permet d'identifier des utilisateurs non suivis jusque-là, mais potentiellement intéressants (les différents liens sont évalués avec un algorithme semblable à l'algorithme du Page Rank).

Les approches exploitant la régression visent à identifier une relation fonctionnelle entre les valeurs en entrée et la sortie souhaitée pour l'algorithme. Plusieurs types de régression existent, et ils sont définis par le type de relation fonctionnelle à établir entre les entrées et la sortie. S'il s'agit d'une fonction linéaire, on va parler de régression linéaire, et ainsi de suite. La régression est l'approche à privilégier lorsqu'on cherche à prédire une valeur continue, comme par exemple prédire une note de l'utilisateur selon les notes qui lui ont été précédemment affectées (Kwon & Kim 2009). (Kim et al. 2005) présentent une étude des performances des différents algorithmes possibles à appliquer lors de recommandations pour des achats en ligne (notamment les arbres de décisions et les réseaux artificiels de neurones). Dans cette étude, le meilleur score a été atteint par la régression logistique (*logistic regression*).

Les ***approches non-supervisées*** ne nécessitent pas d'ensemble de connaissances pré-annoté. Elles se basent sur des connaissances expert ou métier pour faire des déductions ou évaluer les informations disponibles, souvent en employant des mesures statistiques.

Les approches à base de clusters (*clustering*), ou approches de regroupement, reposent sur l'utilisation de mesures statistiques internes pour créer des partitions au sein d'un ensemble de données. Dans un contexte profilage utilisateur, le regroupement peut être utilisé pour déterminer des motifs stéréotypés dans le comportement de l'utilisateur ou encore pour regrouper des utilisateurs dans des sous-groupes homogènes en fonction des fonctionnalités surveillées.

Les réseaux artificiels de neurones décrivent une famille d'approches d'apprentissage statistique visant à imiter la structure synaptique du cerveau. Le réseau neuronal est constitué de cellules ou de neurones uniques. La première couche de neurones, les neurones d'entrée, est activée sur la base des données saisies par le système. Typiquement l'apparition d'un concept recherché au sein d'une page Web va provoquer une activation positive dans un neurone d'entrée spécifique. Cette activation est ensuite propagée vers les couches profondes du réseau de neurones, sur la base d'une fonction définie par l'ingénieur système. Le résultat est déterminé par l'activation atteignant la dernière couche de neurones, les neurones de sortie. Les réseaux de neurones peuvent être appliqués de deux manières différentes, à savoir supervisés et non supervisés. Un exemple d'application non supervisée est constitué par (Martín-Guerrero et al. 2007) qui applique les réseaux artificiels de neurones au regroupement de centres d'intérêts utilisateur. Les auteurs basent leur approche sur la théorie de la résonance adaptative (*Adaptive Resonance Theory*) de (Carpenter & Grossberg 1987), car elle présente l'avantage d'être hautement adaptable lorsque de nouveaux modèles d'utilisation arrivent (ne perturbe pas les regroupements de centres d'intérêts faits précédemment).

Les approches à base de règles se basent sur des connaissances expert, exprimées à travers des règles de la forme (Antécédent -> Conséquent). Le corps de la règle comprend un ensemble de conditions qui doivent être remplies pour déclencher l'exécution de la règle, c'est-à-dire pour que le contenu de la queue de la règle soit déduit et ajouté à la base de connaissances. Ces approches sont utilisées depuis longtemps pour le profilage d'utilisateurs. Le système Grundy, présenté dans (Rich 1979), utilise des règles pour décrire l'appartenance d'un utilisateur à un groupe ou segment donné. L'approche est appliquée dans un contexte de recommandation d'œuvres littéraires. D'autres approches comme (Golemati et al. 2007) utilisent de telles règles pour déterminer le degré d'appariement entre un profil utilisateur et des stéréotypes. (Skillen et al. 2014) reposent leur approche sur une ontologie au sein de laquelle ils intègrent les différentes règles permettant de combiner des attributs individuels du profil utilisateur avec des éléments contextuels (contexte environnemental et applicatif) et ainsi d'adapter la sortie du service considéré par rapport à l'utilisateur.

Le Tableau 12 présente un résumé des propriétés des algorithmes non-supervisés et supervisés, sur la base de ces critères. Pour chacun des algorithmes, on retrouve:

> l'hypothèse de base introduite dans le processus de profilage lors de l'application de l'approche considérée
> les propriétés ou contraintes qui s'appliquent aux données considérées en entrée
> les mécanismes disponibles pour adapter de manière itérative le modèle de profil ainsi généré

La complexité du processus de construction du modèle dépend du degré de complexité avec lequel les instances sont décrites, de la complexité de la fonction de distance et, surtout, du degré d'implication humaine exigé par le processus. Notamment, seule une fraction des algorithmes permet une adaptation itérative, dans le temps, du profil généré. Dans de nombreux cas, le modèle entier doit être recalculé pour intégrer les changements. Dans le cas des règles expert, les modifications ne peuvent être intégrées que manuellement en adaptant le jeu de règles.

Approche	Hypothèse de base	Contraintes	Adaptation du modèle
Classification naïve bayésienne	Instances événements statistiquement indépendantes	Calcul d'une relation statistique entre variables d'entrée et de sortie	Réapprentissage
Réseaux bayésiens	Instances événements statistiquement dépendantes	Construction d'une structure graphe et d'une mesure d'évaluation	Réapprentissage
Machines à vecteurs	Instances séparables d'un point de vue linéaire (utilisant la fonction noyau)	Fonctions noyau appropriées	Réapprentissage

Règles d'association	Description à base de règles des relations possibles	Variables d'entrée discrètes	Adaptation itérative de la base de règles
Algorithmes évolutionnaires	-	Fonction de qualité (*fitness function*)	Réapprentissage
Analyse de liens	Relations sous forme de graphe	Construction de graphe ; Mesure pour évaluer la qualité de la structure des liens	Adaptation itérative de la structure des liens
Exploitant la régression	Relations fonctionnelles entre variables d'entrée et de sortie	Variables continues ; hypothèses concernant les relations fonctionnelles entre variables d'entrée et de sortie	Réapprentissage
Algorithmes supervisés	**Les données sont associées à une classe**	**Apprentissage et corpus d'apprentissage requis**	**Réapprentissage de la classification, souvent incrémental**
Approche	**Hypothèse de base**	**Contraintes**	**Adaptation du modèle**
Clusters	Forme du cluster (dépend de l'approche)	Fonction de distance	Approches itératives et non-itératives
Réseaux de neurones	-	Fonction de distance	Approches itératives et non-itératives
Règles	Description à base de règles des relations possibles	Nécessite des experts connaissant un langage de règles	Aucune, manuellement
Algorithmes non-supervisés	**Regrouper les instances de manière significative**	**Corpus suffisamment grand d'instance exemples**	**Re-calcul du modèle**

Tableau 12.Comparaison des approches de représentation de l'information. (Hoppe 2016)

Le choix de notre approche ne s'est toutefois pas fait uniquement sur la base des considérations ci-dessous. Nous avons complété cette étude des algorithmes supervisés et non-supervisés par une étude, plus générale, portant sur des aspects "conception du système", à savoir: a) comment le choix de la mesure de distance impacte les résultats du système et b) les possibilités de diversification des résultats du profilage.

En effet, dans un problème de profilage, le choix de l'algorithme approprié dépend de plusieurs facteurs. Notamment, le jeu de données disponible et la manière dont les instances qu'il contient sont décrites, influent fortement sur l'algorithme à utiliser. Or, l'hypothèse de base dans tout traitement de données est que les instances décrivent réellement une relation intéressante (qui a du sens, qui est pertinente pour le contexte d'application envisagé). Les algorithmes supervisés nécessitent un corpus pré-annoté pour l'apprentissage. Bien que les algorithmes non supervisés n'exigent pas que les annotations soient connues à l'avance, certains d'entre eux nécessitent encore un ensemble de données initial suffisamment grand (volumineux) pour calculer un modèle d'organisation significatif (sauf les systèmes à base de règles expertes qui ne sont pas impactés par la taille du jeu de données). Les variables décrivant chacune des instances de données peuvent prendre différentes formes - valeurs numériques continues comme discrètes, ou encore annotations par catégories. L'algorithme choisi doit pouvoir gérer le type de variable considéré.

3.2.3.4 Mise à jour d'un profil utilisateur

La plupart des travaux examinés se concentrent sur la construction initiale du profil utilisateur. Ils ont tendance à analyser les données d'entrée disponibles sous la forme d'un lot (ou d'un ensemble) et à stocker les résultats. Cette approche est difficilement implémentable (ou adaptable) dans l'environnement

Web d'aujourd'hui. D'une part, les utilisateurs parcourent une multitude de pages chaque jour. Un traitement par lots nécessite de retraiter l'ensemble des pages vues par le passé - conduisant à des quantités de données de plus en plus importantes et accumulées à chaque clic. D'autre part, les contenus profilés doivent s'adapter de manière dynamique, notamment à des modifications et des évolutions pouvant survenir dans les données en entrée. Il est malheureusement impossible de satisfaire cette exigence si chaque mise à jour requiert une analyse complète de l'ensemble des informations dont nous disposons concernant l'utilisateur.

La mise à jour d'un profil utilisateur devient un problème important dans le cas des systèmes devant gérer des données massives, dans des environnements hautement dynamiques (comme c'est le cas pour le système MindMinings). Toutefois, l'état de l'art du domaine a montré, de manière surprenante, que peu de travaux de recherche s'attaquent à ce problème. Dans ce qui suit, je présente les principales approches identifiées.

La mise à jour d'un profil utilisateur peut comporter souvent deux principales étapes:

> ***Étape 1 - L'intégration de nouvelles informations dans le profil existant et l'adaptation, le cas échéant, des pondérations et des systèmes de pondération***
Cette étape est souvent simple lorsqu'implémentée de manière binaire e.g. les nouveaux contenus sont directement ajoutés à la structure de profil existante. Cela devient plus complexe lors de la mise à jour de structures de profil pondérées. En effet, l'adaptation correcte et la normalisation des pondérations exigent un cadre mathématique bien défini. (Carreira et al. 2004) donnent une solide justification de cette affirmation. Ils présentent une approche utilisant des retours de l'utilisateur (positifs ou négatifs) afin d'augmenter (respectivement diminuer) d'un certain pourcentage la pondération associée à un article. Comme discuté par les auteurs, ce type d'approche permet une adaptation rapide. Dans (Vu et al. 2015), les auteurs proposent une approche adressant le problème des informations abondantes en utilisant des échantillons pour calculer le profil utilisateur. Ils construisent trois profils d'utilisateurs, à long terme, quotidiens et axés sur la session. Chacun d'eux est obtenu en analysant un nombre fixe de documents pertinents, sur la période correspondant aux sessions considérées. L'algorithme d'activation diffuse (*Spreading Activation* ou SA) est une approche spécifiquement adoptée pour les profils basés sur des structures de graphe. Cette approche permet de mettre à jour les pondérations associées à chacun des concepts d'intérêt en exploitant la structure graphe sous-jacente. Plusieurs exemples d'implémentation peuvent être trouvés dans la littérature, allant d'une version basique de l'algorithme (Sieg et al. 2007) à une version dérivée contrainte (*Constrained SA*) comme dans (Vallet et al. 2007).
> ***Étape 2 - La suppression des informations obsolètes pour conserver le profil à une taille raisonnable***
Les fonctions de décroissance du temps (*time decay functions*) représentent une version plus sophistiquée des approches par "fenêtre glissante" (*sliding window*). Au lieu de supprimer les données lorsqu'elles atteignent un âge donné, l'importance des données vieillissantes est "lissée". La fonction de décroissance est intégrée dans le calcul du résultat final. Cela diminue l'influence des données plus anciennes sur le résultat et favorise le contenu informationnel associé aux activités plus récentes de l'utilisateur. Ce type d'approche suppose que l'attention de l'utilisateur varie avec le temps. Certains intérêts de base peuvent perdre de leur importance avec le temps. Ainsi, les informations récentes devraient recevoir une plus grande attention lorsqu'évaluées. Ce type d'approche a été appliqué à des profils utilisateurs reposant sur différents types de représentations: à base de mots-clés (Teevan et al. 2005), à base de réseaux sémantiques (Gasparetti & Micarelli 2005), à base de hiérarchies de concepts (Li et al. 2014), ou encore concepts d'ontologie (Vallet et al. 2007).

En guise d'alternative, le système de profilage peut s'appuyer sur des mesures basées sur le contenu pour décider quelles informations supprimer ou conserver. Deux approches sont discutées dans (Bouchachia et al. 2014): chaque fois que les données entrantes sont trop différentes du profil utilisateur, un nouveau vecteur d'intérêts est construit. Si le système n'a pas encore atteint le nombre maximal de vecteurs profils, le nouvel élément est simplement ajouté à la base de vecteurs. Sinon, il est soit fusionné avec un vecteur existant suffisamment similaire, soit le vecteur existant le plus différent est rejeté comme étant obsolète. D'autres approches exploitent principalement les techniques d'agrégation d'informations.

Pas toutes les procédures de mise à jour gèrent ces deux étapes. Si le profil utilisateur est constitué d'un ensemble de pondérations associées à un ensemble fixe de catégories, il n'est pas utile de supprimer du contenu (2e étape). Cependant, l'adaptation progressive des pondérations peut exiger une fonction de

pondération sophistiquée. Le système peut appliquer une approche de type "fenêtre glissante" (sliding window) permettant d'éliminer les données ayant une certaine ancienneté dans le système. Si la fenêtre de temps choisie est suffisamment réduite, une approche de traitement par lots réguliers peut fournir des résultats dynamiques en temps rapide.

Toutefois, les approches décrites ci-dessous influencent uniquement le contenu du profil. Elles peuvent, par exemple, ignorer un intérêt spécifique de l'utilisateur si celui-ci n'est pas supporté par les données en entrée. Il n'y a cependant aucune méthode qui influence l'interprétation abstraite des données d'entrée elles-mêmes. Une telle approche permettrait d'associer une nouvelle pondération en différenciant des critères d'évaluation par type de canal d'entrée utilisé par les données d'entrée. Typiquement cela permettrait de faire la distinction entre des données fournies en entrée de manière explicite par l'utilisateur et des données fournies de manière implicite.

3.2.3.5 Évaluation d'un profil utilisateur

L'étape d'évaluation d'un profil utilisateur est aussi importante que sa conception. L'évaluation de la qualité permet la comparaison avec d'autres approches ainsi la quantification du degré d'amélioration obtenu grâce à la nouvelle approche. En outre, l'évaluation permet de découvrir les faiblesses de la nouvelle approche et de trouver des moyens de les corriger. Étant données ces considérations, nous pourrions nous imaginer qu'une multitude de plateformes d'évaluation standard sont disponibles. La réalité est malheureusement différente. De nombreuses approches délèguent cette étape aux travaux futurs (Adar et al. 1999), (Parent et al. 2001), (McGowan et al. 2002), (Grcar et al. 2005) ou l'omettent tout simplement (Lieberman 1995), (Horvitz et al. 1998), (Von Hessling et al. 2005), (Stan et al. 2008), (Confalonieri et al. 2012) . Dans ce qui suit, je vais discuter les difficultés associées à l'évaluation de profils qui pourraient expliquer ce constat. J'examine ensuite quelques approches d'évaluation de systèmes de profilage.

Certains travaux font des comparaisons par rapport à des bases aléatoires (Rich 1979), (Teevan et al. 2005), alors que d'autres comparent leurs approches à des systèmes existants, non-personnalisés (Teevan et al. 2005), (Sieg et al. 2007), (Vallet et al. 2007) ou bien personnalisés, ciblant la même tâche (Teevan et al. 2005), (Eyharabide & Amandi 2012), (Mylonas et al. 2008) ou encore à différentes paramétrisations du même système (Sakagami & Kamba 1997), (Chu & Park 2009), (Sela et al. 2015).

Pour évaluer leurs approches, de nombreux auteurs adoptent deux mesures standard issues du domaine de la recherche d'informations: la précision et le rappel. Les systèmes conçus dans ce contexte visent à aider les utilisateurs lors de la recherche d'informations. Lorsqu'une évaluation d'un système de profilage utilisateur est effectuée, les auteurs mettent souvent l'accent sur la discussion des résultats. Les méthodes de test sont rarement décrites en détail, de même que les propriétés des sujets de test potentiellement utilisés, ou encore les limites des mesures quantitatives adoptées.

Globalement, on peut séparer les techniques d'évaluation de système de profilage en deux groupes principaux:

> Celles qui utilisent les applications hôtes pour quantifier la qualité du profil,
> Celles qui évaluent directement le résultat du processus de profilage.

Dans le premier cas, les chercheurs ont choisi un environnement applicatif spécifique qui bénéficie de la personnalisation - et comparent (i) les différences de performances entre un système personnalisé et un système non personnalisé, (ii) les performances des différentes approches de personnalisation, (iii) différentes configurations pour leur propre approche de profilage. Ce type d'approche permet d'adopter des mesures d'évaluation communes au domaine d'application considéré. Cependant, le transfert vers une application arbitraire peut modifier les résultats de l'évaluation.

Les approches d'évaluation directe confrontent les individus avec le modèle de profil utilisateur construit et recueillent leurs commentaires. Bien que cela semble être une approche intuitive, elle a ses limites. Comme indiqué dans (Wærn 2004), les utilisateurs ne comprennent pas toujours et ne jugent pas correctement le contenu du profil - leurs remarques pourraient donc ne pas donner une évaluation précise (voire même correcte) du résultat du profilage. De plus, cette approche individualiste demande beaucoup de travail, car elle implique d'organiser des interviews avec chaque utilisateur du système. En conséquence, ce type d'évaluation est difficilement applicable à une grande échelle.

Les approches citées ci-dessus s'adressent à des applications du domaine de la recherche personnalisée d'informations ou celui du commerce électronique. Les auteurs proposent des évaluations basées sur:

a) des ensembles de données thématiques de référence (*benchmarks*) e.g. (Vallet et al. 2007), (Mylonas et al. 2008).
b) des données générées par l'utilisateur e.g. (Sakagami & Kamba 1997), (Eyharabide & Amandi 2012), (Rosaci & Sarnè 2014)
c) des données générées artificielles qui simulent le comportement humain (Moukas 1997), (Román & Velásquez 2014).

L'utilisation de données simulées a l'avantage d'être indépendante des utilisateurs réels, et donc de s'absoudre des problèmes liés à la confidentialité des données. Malheureusement, la génération de données utilisateur réalistes demeure une question de recherche ouverte.

Les ensembles de données de référence (*benchmarks*) offrent une base de données normalisée qui permet de reproduire et de comparer les résultats des tests. Cependant, la maintenance de tels jeux de données est une tâche fastidieuse. En conséquence, les entités responsables de la publication de tels ensembles de données arrêtent de les suivre et de les mettre à jour, une fois publiés. De plus, les ensembles de données de référence sont souvent spécifiques à une application et les résultats ne peuvent pas être généralisés à d'autres applications, même si elles sont dans un même domaine.

L'utilisation de données réelles générées par l'utilisateur conduit à un inventaire de données réaliste. Cependant, la collecte et le stockage de telles données impliquent souvent des considérations en termes de confidentialité. En effet, dans la plupart des cas cités ci-dessus, il n'a pas été possible de publier des données réelles sur les utilisateurs dans le document d'évaluation de l'approche. En conséquence, les résultats des tests ne peuvent pas être reproduits par d'autres chercheurs, les performances du système ne peuvent être comparées à celles d'autres prototypes. Un autre problème, qui est rarement discuté dans les publications, est le biais de la base d'utilisateurs sous-jacente. Souvent, les chercheurs utilisent des traces collectées sur un serveur Web universitaire - ce qui permet de fixer certaines hypothèses sur la population observée d'utilisateurs.

3.2.4 Approche et résultats

3.2.4.1 Approche

Les travaux de recherche présentés ici traitent de la conception d'un nouveau type de système de profilage utilisateur. La conception de tels systèmes dépend dans une large mesure du contexte d'application environnant, qui, dans le cadre cette thèse, fut le domaine de la publicité numérique. Sur la base de l'état de l'art précédent, les paragraphes suivants présentent les choix de conception pour le système MindMinings ainsi que les défis associés, et ce pour chacune des phases identifiées dans le processus de profilage.

Le choix d'une technique pour la collecte d'informations a été fait sur la base de critères spécifiques à l'application. Ceux-ci ont été extraits à partir d'une étude des techniques que l'entreprise souhaitait utiliser dans l'application de profilage. En effet, reposant sur des techniques collaboratives et des profils utilisateurs basés sur des stéréotypes, le système doit pouvoir interconnecter les profils des différents utilisateurs entre eux. Ceci est impossible à réaliser dans un contexte où le profil utilisateur est uniquement stocké côté client. Basé sur une collecte d'informations implicites, le système MindMinings exploite, en tant que principales sources d'informations pour le processus de profilage, les journaux de navigation Web, collectés par les fournisseurs d'accès en ligne côté serveur. Ceux-ci contiennent généralement des informations de base sur le contexte technologique de l'utilisateur (OS, version, moteur d'affichage, etc.) et l'identifiant du contenu Web demandé. Ceci représente le point de départ dans la construction du profil utilisateur. Des techniques d'analyse supplémentaires sont implémentées pour déduire des caractéristiques de profil de haut niveau. Pour ce faire, le système de profilage MindMinings se concentre sur la modélisation sémantique des ressources Web. En conséquence, le flux de travail du système intègre des techniques d'extraction Web et de traitement du langage naturel.

Au niveau de la représentation d'informations utilisateur, l'état de l'art a montré une tendance générale dans les travaux en cours en 2012 à utiliser des techniques de modélisation des connaissances de plus

en plus sophistiquées. En effet, les approches les plus récentes à l'époque avaient fait le choix des ontologies pour la structurer les connaissances du système. L'état de l'art a aussi montré, qu'en 2012, ces approches n'avaient pas encore été implémentées à une échelle industrielle. Pour combler cette lacune, nous avons fait le choix pour le système MindMinings d'utiliser une ontologie comme stockage central pour toutes les informations relatives aux profils. La collaboration avec l'entreprise partenaire nous a permis de concevoir et de tester le prototype en fonction de critères industriels. Les **défis à résoudre** par rapport à ce choix de conception sont nombreux:

> **Consensus** - La conception de l'ontologie s'est fait en coopération étroite avec des experts du domaine (échanges par écrit, réunions régulières en face à face) et en suivant le processus itératif décrit dans (Noy & McGuiness 2001). Ceci a suscité des discussions intenses non seulement entre l'équipe d'ingénierie de l'ontologie et les employés de l'entreprise, mais également entre les experts du domaine eux-mêmes. Ces différents échanges nous ont permis d'identifier les principaux concepts du domaine ainsi que les relations entre eux. Le modèle ontologique présenté dans la section suivante illustre le consensus atteint.
> **Modularité** - Afin de maximiser la compréhension des éléments de l'ontologie, ceux-ci ont été structurés de manière modulaire. Chaque module de l'ontologie permet de capturer un aspect sémantique du processus de profilage. Des relations sémantiques relient les différents modules entre eux.
> **Personnalisation** - Une caractéristique centrale du modèle d'ontologie présenté est de se focaliser strictement sur le domaine d'application. Le système de profilage doit rester solide dans un environnement à forte intensité de données. Il doit donc intégrer, traiter et évaluer de grandes quantités de données dans des délais raisonnables. Pour ce faire, il est important de limiter la base d'informations dans le système au strict nécessaire et d'éliminer tous les éléments qui ne concernent pas le contexte d'application spécifique.

Concernant l'exploitation des informations d'un profil utilisateur, l'état de l'art a permis d'identifier deux principales catégories d'approches:

> celles basées sur les connaissances et utilisant des modèles de connaissances experts métiers pour déduire, à partir des données en entrée, les attributs finaux du profil;
> celles basées sur les données et utilisant des mesures statistiques sur les données en entrée ou des corpus d'apprentissage.

Composer un profil présentant des caractéristiques intéressantes dépend fortement du contexte de la campagne publicitaire concernée. En effet, les segments marketing, identifiant les profils "intéressants" pour un produit donné, changent d'une campagne à l'autre, d'un produit à l'autre. Ces segments sont définis par des experts du domaine de la publicité en ligne, à partir de leur expérience et connaissances. Nous avons donc choisi de baser le processus de profilage du système MindMinings sur une ***approche exploitant les connaissances***. L'avantage est double: d'une part cela permet d'intégrer des expertises métier, et d'autre part cela élimine le besoin de constituer un corpus d'apprentissage pour chaque nouvelle campagne publicitaire.

Par rapport à ce choix de conception, le principal défi concerne la ***représentation de la quantification du degré d'appartenance d'un utilisateur à un segment marketing***. En effet, il est difficilement imaginable d'avoir des profils utilisateurs correspondant en tout point (recouvrement total) aux contraintes composant un tel segment. Suite à l'étude des informations extraites à partir de journaux de navigation Web d'utilisateurs, elles ne permettent pas de tirer des conclusions claires quant aux caractéristiques des profils utilisateurs sous-jacents. Un exemple de problème serait par exemple de déterminer le sexe d'un utilisateur sur la base de sa navigation Web. Or, c'est l'hypothèse à la base de ces travaux de recherche - l'analyse des pages Web consultées par l'utilisateur permet de déduire des caractéristiques de son profil. Afin d'extraire des connaissances utiles à partir des journaux de navigation Web fournis par l'entreprise, nous avons donc fait l'hypothèse que les attributs de profil ainsi déduits seraient pondérés afin de capturer le degré d'appartenance d'un profil utilisateur par rapport à un segment marketing. Au niveau de l'ontologie conçue, cela soulève le problème de la gestion de l'incertitude avec des technologies sémantiques.

Enfin, concernant la mise à jour des profils utilisateurs, la plupart des approches étudiées se focalisaient sur une composition initiale correcte du profil utilisateur. Les différents mécanismes de mise à jour

comprennent l'intégration / la suppression d'éléments du profil et l'adaptation d'éventuelles valeurs de scores ou de pondérations. Ayant fait le choix de baser le système MindMinings sur des connaissances d'experts pour la composition et la déduction de caractéristiques d'un profil, la ***mise à jour d'un profil se fera en adaptant les règles à utiliser pour la déduction selon les besoins de la campagne concernée***. Le système n'utilisera pas une procédure de mise à jour automatique. Les experts doivent pouvoir modifier et créer de nouveaux segments marketing, et les connaissances de l'ontologie sont automatiquement classées selon ces segments (calcul automatique du degré d'appartenance d'un profil utilisateur existant par rapport à un segment modifié ou nouvellement défini). Ceci soulève la question (centrale à l'ensemble des travaux de recherche présentés ici) de l'équilibre à atteindre entre l'expressivité et la décidabilité du modèle ontologique. D'après l'état de l'art des langages de description d'ontologies disponible, OWL-DL représente le meilleur choix en vue d'établir un tel équilibre. Toutefois, l'expressivité d'OWL-DL ne suffit pas à capturer l'ensemble des subtilités du processus de profilage. Les règles logiques composant les segments marketing représentent une extension du modèle ontologique de base. Basées sur le langage SWRL, ces règles ont l'avantage d'utiliser une syntaxe conjonctive facilement compréhensible et interprétable, même pour des personnes non spécialistes de l'ingénierie des ontologies. Il est par contre essentiel de fournir aux différents experts des outils intuitifs (e.g. interfaces graphiques) pour spécifier leurs connaissances et éditer les différentes règles et segments.

Les différents choix de conception présentés ci-dessus ont permis les **contributions** suivantes:

- **Conception d'une ontologie du domaine de la publicité en ligne adaptée pour le profilage d'utilisateurs** - Au niveau de la représentation des informations d'un profil utilisateur, le système MindMinings repose sur une ontologie propre, l'ontologie du profil utilisateur. L'état de l'art du domaine a montré qu'un tel modèle ontologique n'existait pas, soit parce que les approches précédentes n'étaient plus maintenues, soit parce qu'aucun standard n'a été défini pour le domaine concerné. Par conséquent, nous avons dû développer un modèle ontologique pour permettre au système MindMinings de capturer et interpréter les informations de profil nécessaires.
- **Prise de compte de l'incertitude** - L'exploitation des connaissances a été faite selon un modèle du processus de profilage observé par l'entreprise. Cela a donné naissance à une extension du modèle ontologique de base afin de permettre la gestion d'informations incertaines (ou avec un degré de certitude)
- **Conception d'interfaces intuitives** - La mise à jour des différents profils utilisateurs exploite les règles logiques utilisées dans la définition des segments marketing. Des interfaces ont été spécifiquement développées afin de permettre une définition aisée de règles logiques (en glissant-déposant des concepts de l'ontologie et en plaçant des opérateurs logiques entre ceux-ci).

Ces contributions sont présentées et discutées en deux parties: dans un premier temps, je présente les contributions à un niveau conceptuel (modèle ontologique de base, gestion de l'incertitude), puis dans un deuxième temps, je présente l'implémentation du modèle conceptuel, à travers le prototype développé pour le système MindMinings.

3.2.4.2 Contributions scientifiques

Pour chacune des étapes identifiées pour le processus de profilage, nos contributions sont résumées ci-dessous:

- ***Collection des informations*** - Les données utilisées en entrée de notre prototype ont été fournies par l'entreprise, sous la forme de journaux de navigation Web d'utilisateurs réels. Devant la multitude de techniques permettant d'extraire du contenu à partir de pages Web, nous avons été dans l'impossibilité de les intégrer toutes au sein d'un même prototype. Notre représentation des ressources Web se focalise sur leurs éléments textuels. De manière similaire aux principales techniques d'indexation, notre système utilise une approche statistique pour identifier les mots-clés les plus importants dans une page Web. Une fois ces mots-clés extraits, nous utilisons une plateforme complémentaire et existante pour identifier les entités nommées apparaissant dans le texte considéré. Les termes ainsi découverts sont ensuite qualifiés sémantiquement à l'aide de base de connaissances telles DBpedia. Cette qualification sémantique permet d'associer une URI à chaque mot-clé extrait initialement et donc de constituer une description avec des concepts sémantique du domaine de discours de la ressource Web considérée. Ainsi, pour chaque page Web dans l'historique de navigation de l'utilisateur, nous avons un ensemble fini de concepts sémantiques, que nous utilisons

afin de déterminer "de quoi traite la page Web". Notre approche permet donc de "classer" les pages du Web par rapport à une taxonomie de catégories (ou d'univers), à la construction de laquelle j'ai fortement participé (étant donné mes travaux de thèse sur les classifications de services et de biens e.g. NAICS).

> ***Représentation des informations*** - Toutes les informations sont stockées dans la base de connaissances (ou l'ontologie), élément central du système MindMinings (choix de conception justifié précédemment). L'ontologie MindMinings repose sur une terminologie clairement définie, développée en collaboration avec des experts du domaine. Afin d'éviter les incompréhensions et erreurs d'interprétation, chaque concept possède une description en langage naturel. A chaque fois que cela a été possible, la description des concepts de l'ontologie a été étendue avec des références vers des référentiels externes de connaissances e.g. DBpedia. Pour ce faire, des relations sémantiques (e.g. **`owl:sameAs`**, **`owl:equivalentClass`**, **`owl:equivalentProperty`**) ont été utilisées. L'ontologie présente l'avantage d'être facilement extensible pour s'adapter aux caractéristiques des utilisateurs (possibilité d'ajouter des concepts) et aux évolutions des segments marketing (possibilité de modifier l'ensemble de règles de déduction pour chaque segment, de même que possibilité d'ajouter de nouveaux segments). De plus, d'éventuelles incohérences introduites, par exemple, lors de l'ajout de nouveaux segments peuvent être détectées en utilisant un raisonneur au-dessus de la base de connaissances.
> ***Exploitation du profil*** - Cette étape exploite les connaissances obtenues à partir des informations en entrée et produit le profil final, autrement dit déduit les attributs finaux du profil sur la base des connaissances et règles contenues dans l'ontologie. L'état de l'art réalisé a permis d'identifier bon nombre d'algorithmes supervisés comme non-supervisés. Contrairement aux approches trouvées dans la littérature, dans le contexte du projet MindMinings, l'objectif est d'internaliser l'exploitation des profils construits. En ce sens, notre prototype ne permet pas l'export du profil construit vers d'autres représentations (éventuellement moins expressives ou moins flexibles). De plus, au lieu de se référer à des mécanismes externes pour exploiter les informations contenues dans un profil, le système MindMinings utilise des mécanismes internes pour déduire la valeur finale des attributs du profil. Notre choix de formalisme (e.g. les règles logiques) se justifie car, d'une part, il est suffisamment expressif pour capturer la complexité du processus de profilage et, d'autre part, les raisonneurs existants peuvent interpréter les connaissances utilisées.
> ***Mise à jour du profil*** - Dans le prototype développé pour le système MindMinings, le processus de mise à jour est implémenté au travers des règles de déduction conçues par les experts du domaine. Dans cette version du prototype, la mise à jour des règles sous-jacentes incombe aux experts du domaine. Nous avons tout de même offert un support en vue d'un tel fonctionnement, notamment à travers la définition d'interfaces homme-machine (IHM) pour l'édition, l'ajout et la suppression des règles logiques constituant la base de règles de l'application.

3.2.4.2.1 Collection d'informations - Ontologie

L'ontologie représentant un profil utilisateur dans le contexte du projet MindMinings comporte deux parties, décrites dans les paragraphes suivants:

> **L'ontologie de base** avec les classes, les relations entre ces classes ainsi que les contraintes s'appliquant sur ces classe et relations, modélisant les connaissances issues du processus de profilage (partant des données de navigation brutes et allant jusqu'à des caractéristiques profil de haut niveau)
> **La couche logique** comprenant les règles logiques permettant de "donner vie" aux classes et relations de l'ontologie de base. L'idée est d'étendre l'ontologie par des concepts et des règles afin que le raisonneur puisse automatiquement déduire des informations supplémentaires sur l'utilisateur profilé.

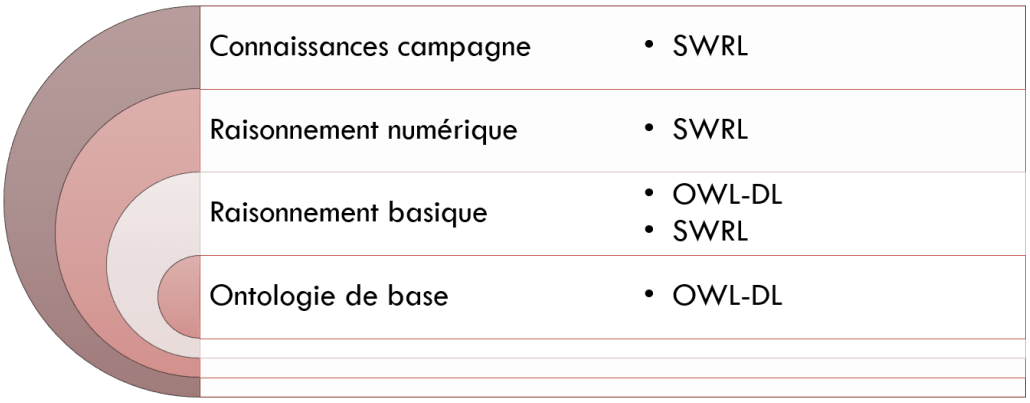


Figure 21. Niveaux de raisonnement implémentés dans l'ontologie MindMinings (Hoppe 2016)

Comme précisé avant, nous avons souhaité concevoir une ontologie modulaire afin faciliter son interprétation ainsi que son implémentation par les employés de l'entreprise (à terme les utilisateurs finaux du système MindMinings). Pour ce faire, nous avons considéré les modules suivants, tels qu'illustrés par la Figure 22:

> Le module "Journal de navigation" contient les concepts capturant l'historique de navigation de l'utilisateur
> Le module "Qualification des ressources Web" contient les concepts correspondant aux ressources Web et leur classification en des catégories de sujets.
> Le module "Univers" contient la taxonomie de catégories conçue
> Le module "Profil utilisateur et segments" capture à la fois les attributs de base du profil utilisateur et les segments marketing haut-niveau auxquels le profil appartient (selon un degré d'appartenance)

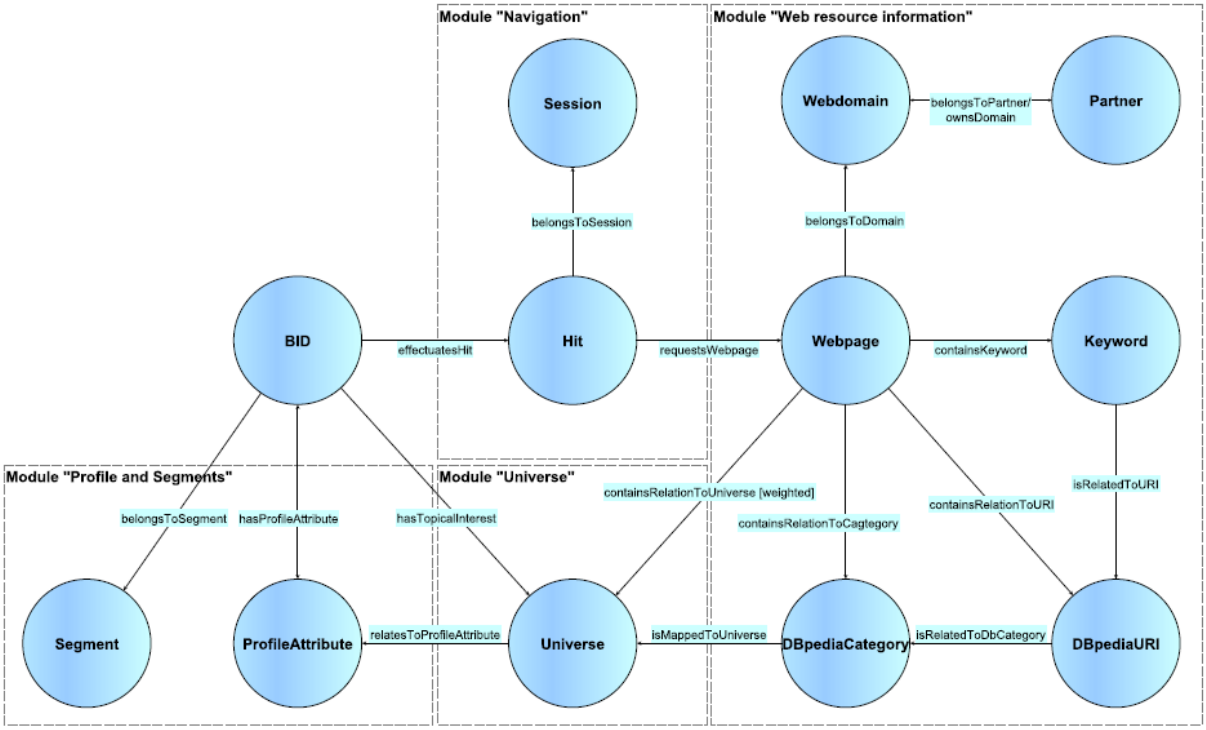


Figure 22. Vue schématique des différents modules composant l'ontologie MindMinings (Hoppe 2016)

Un concept central de l'ontologie est le concept représentant l'utilisateur proprement dit. Afin de respecter la terminologie utilisée par les experts du domaine, ce concept s'appelle **`BID`** (pour *Browser Identification*) et est placé en dehors de tout module. Ceci est justifié par le fait que l'entreprise utilisait une approche à base de cookies pour identifier les utilisateurs profilés au travers de plusieurs plateformes Web (les problèmes liés à l'identification d'utilisateurs au travers de cookies ont été discutés et étudiés durant la thèse, et ont été considérés comme hors sujet par rapport aux travaux de recherche sous-jacents).

Dans l'ontologie MindMinings, un utilisateur est donc identifié au travers d'une instance du concept **BID**, plus particulièrement à travers sa propriété de type de données **hasBID**, permettant de stocker la chaîne de caractères (**xsd:string**) correspondant à son identifiant unique dans le système. Les instances **BID** contiennent une autre propriété de type de données (**hasNumberOfHits**) qui quantifie le nombre de clics de l'utilisateur sur des pages Web. Enfin le lien entre un utilisateur et des univers est modélisé au travers de trois propriétés: la propriété **hasTopicalInterest** pour spécifier qu'un utilisateur donné a un intérêt par rapport à un domaine donné, les propriétés **hasProfileAttribute** et **belongsToSegment** font le lien entre des attributs bas niveau du profil et des segments marketing haut-niveau. L'équation 1 résume les éléments décrits plus haut concernant le concept de **BID**.

BID	$\sqsubseteq$	$\top$	Equation 1
	$\sqcap$	$\exists hasBID.(xsd:string)$	
	$\sqcap$	$\exists hasNumberOfHits.(xsd:integer)$	
	$\sqcap$	$\leq 1\ hasEnd.(xsd:dateTime)$	
	$\sqcap$	$\exists\ effectuatesHit.Hit$	
$\exists hasTopicalInterest.\top$	$\sqsubseteq$	BID	
$\exists hasProfileAttribute.\top$	$\sqsubseteq$	BID	
$\exists belongsToSegment.\top$	$\sqsubseteq$	BID	

Le module "***Journal de navigation***" comprend deux principaux concepts à savoir: **Hit** et **Session**. Chaque instance de type **Hit** regroupe les informations disponibles pour une interaction utilisateur individuelle. Une telle instance capture l'ensemble des données contenues sur une ligne d'un journal de navigation Web. Selon les différentes informations contenues dans le journal de navigation, les propriétés de type de données de la classe **Hit** permettent de capturer les contextes suivants d'un clic sur une page Web:

- temporel (extrait à partir de l'étiquette temporelle accompagnant chaque entrée du journal de navigation) comme l'heure dans la journée, le jour de la semaine, etc.
- technologique (extrait à partir de la chaîne de l'User Agent) comme le type de terminal ou le système d'exploitation utilisé)
- géographique (sous réserve d'informations disponibles) comme le code postal, le pays ou encore la position géographique (latitude/longitude) de l'utilisateur

Le concept **Session** correspond à une session de navigation et regroupe des événements utilisateur en fonction de leur proximité temporelle. En suivant l'interprétation du partenaire industriel, les limites d'une session sont déterminées par une heuristique de base: les différentes instances de **Hit** sont supposées appartenir à la même instance de **Session** si leurs horodatages ne sont pas espacés de plus de trente minutes. Le concept de **Session** possède deux propriétés de type de données, **hasBegin** et **hasEnd**, contenant les étiquettes de temps associées aux instances de **Hit** délimitant la session. La propriété objet **belongsToSession** permet de relier une instance de **Hit** à une instance de **Session**.

$Session$	$\sqsubseteq$	$\top$	Equation 2
	$\sqcap$	$\leq 1\ hasBegin.(xsd:dateTime)$	
	$\sqcap$	$\geq 1\ hasBegin.(xsd:dateTime)$	
	$\sqcap$	$\leq 1\ hasEnd.(xsd:dateTime)$	

Le module "***Qualification des ressources Web***" représente la source d'informations centrale au système MindMinings. Chaque page Web subit un certain nombre d'étapes d'analyse pour (i) découvrir les concepts sémantiques dans son contenu; (ii) la relier à un ensemble de catégories spécifiques à un domaine (univers). Les classes suivantes sont utilisées par ce module afin de stocker les différents résultats intermédiaires du processus de qualification:

La classe **Webpage** regroupe l'ensemble des informations concernant une page Web, et possède différentes propriétés permettant de relier une page Web à un univers donné. Cette connexion entre une page Web et un univers représente la base du processus de profilage implémenté dans MindMinings. Les différentes instances de la classe **Hit** permettent de relier un utilisateur individuel aux ressources Web consultées. Ces ressources Web sont associées à des catégories de sujets (ou univers), à travers la

propriété **containsRelationToUniverse**. Une concaténation des propriétés **effectuatesHit**, **requestsWebpage** et **containsRelationToUniverse** permet de directement connecter des instances de la classe **BID** à des catégories d'intérêt, sur la base de leur historique de navigation.

$Webpage$	$\sqsubseteq$	$\top$	Equation 3
	$\sqcap$	$\leq 1\ hasURL.(xsd:uri)$	
	$\sqcap$	$\geq 1\ hasURL.(xsd:uri)$	
	$\sqcap$	$\forall\ containsRelationToUniverse.Universe$	
	$\sqcap$	$\forall\ hasWebDomain.WebDomain$	

La classe **Webdomain** capture les domaines de haut-niveau associés aux différentes ressources Web analysées. Par rapport à notre contexte d'application, le domaine d'une page Web est identifié au travers d'une combinaison des domaines de premier et de deuxième niveau (sur la base de l'approche officielle DNS (Mockapetris & Dunlap 1988). L'URL d'un domaine Web est stocké à travers la propriété de type de données **hasURL**, qui est fonctionnelle et sert d'identifiant pour un domaine Web. Un domaine Web est relié aux ressources qui en font partie via la propriété **isDomainOf** (qui est l'inverse de la propriété **hasWebDomain**).

$Webdomain$	$\sqsubseteq$	$\top$	Equation 4
	$\sqcap$	$\forall\ isDomainOf.Webpage$	

La classe **Keyword** permet de modéliser les mots-clés extraits à partir des contenus textuels des pages Web. La propriété de type de données **hasLabel** (fonctionnelle) contient le mot-clé original (tel qu'extrait à partir du texte) et permet donc d'identifier le mot-clé dans la base de connaissances. A chaque fois que cela est possible, les différentes instances de la classe **Keyword** sont reliées à des ressources DBpedia équivalentes par le biais de la propriété objet **isRelatedToURI**.

$Keyword$	$\sqsubseteq$	$\top$	Equation 5
	$\sqcap$	$\leq 1\ hasLabel.(xsd:string)$	
	$\sqcap$	$\geq 1\ hasLabel.(xsd:string)$	
	$\sqcap$	$\geq 0\ isRelatedToURI.DBpediaURI$	

La classe **DBpediaURI** permet de caractériser les URI de concepts et relations présents dans DBpedia. L'URI proprement dite de la ressource est stockée grâce à la propriété de type de données **hasURI** (fonctionnelle). Pour chaque ressource ainsi identifiée, on fait le lien avec la catégorie DBpedia à laquelle elle appartient (via la propriété **isRelatedToDBpediaCategory**)

$DBpediaURI$	$\sqsubseteq$	$\top$	Equation 6
	$\sqcap$	$\leq 1\ hasURI.(xsd:uri)$	
	$\sqcap$	$\geq 1\ hasURI.(xsd:uri)$	
	$\sqcap$	$\forall\ isRelatedToDbCategory.DBpediaCategory$	

DBpediaCategory est une classe modélisant des concepts abstraits, de haut niveau, de la base de connaissances DBpedia. Cette classe permet de capturer des catégories génériques, récupérées à travers les URIs DBpedia associées à des ressources Web. Il s'agit d'une sous-classe de la classe DBpediaURI, nécessaire pour exprimer le fait que toute catégorie DBpedia (**DBpediaCategory**) est, avant tout, une URI DBpedia (**DBpediaURI**)

$DBpediaCategory$	$\sqsubseteq$	$DBpediaURI$	Equation 7
	$\sqcap$	$\forall\ isMappedToUniverse.Universe$	

Le module "***Univers***" contient un seul concept de haut niveau, à savoir la classe **Universe**. La taxonomie de catégories construite devient un ensemble de sous-classes de ce concept. Le nom du concept trouve son origine dans la terminologie interne du partenaire industriel. Il a été retenu pour souligner la personnalisation, à travers l'ontologie, de la taxonomie de catégories de sujet. La ***spécialisation du modèle de connaissance par rapport au domaine de la publicité en ligne*** représente une caractéristique importante du

système de profilage MindMinings. Elle permet de limiter l'information en circulation au besoin précis - en profilant l'utilisateur directement à l'aide de fonctions pertinentes sur le plan commercial. Les noms (en langage naturel) des différents univers peuvent s'avérer complexes. Lorsque l'on considère les 26 catégories (initialement définies, étendues par la suite), leurs noms contiennent parfois des espaces et des signes de ponctuation qui ne sont pas autorisés pour une URI. Par conséquent, les noms des instances d'univers sont composés de leurs titres en langage naturel, en supprimant les caractères non supportés dans le schéma standard des URIs. La propriété de type de données **`hasLabel`** (contenant des données de type **`xsd:string`**) est utilisée pour stocker les noms d'univers avec l'ensemble des caractères de ponctuation et espacement. Cette classe possède de nombreuses connexions entrantes, et est bien plus souvent interprétée en tant que portée de propriétés objet que domaine. C'est le cas pour les relations **`containsRelationToUniverse`** (lien entre une ressource Web et un univers) ou encore **`hasTopicalInterest`** (modélisant l'intérêt d'un utilisateur par rapport à un univers donné).

$Universe$	$\sqsubseteq$	$\top$	Equation 8
	$\sqcap$	$\leq 1\ hasLabel.(xsd{:}\ string)$	
	$\sqcap$	$\geq 1\ hasLabel.(xsd{:}\ string)$	
	$\sqcap$	$\forall\ isMappedToCategory.DBpediaCategory$	

Le module ***"Profil utilisateur et segments"*** permet de a) relier des instances de la classe BID à leur profil par rapport à un domaine spécifique, et b) spécifier des connaissances du domaine supplémentaires, sous la forme d'attributs de profil.

La classe **`ProfileAttribute`** comprend différents sous-classes correspondant à une modélisation des éléments de profil définie en collaboration avec le partenaire industriel. Ces éléments se divisent en deux groupes: les attributs socio-démographiques (tels que l'âge ou la catégorie socio-professionnelle de l'utilisateur) et les attributs comportementaux (tels que la langue préférée ou le type de navigation de l'utilisateur).

$ProfileAttribute$	$\sqsubseteq$	$\top$	Equation 9
$SocioDemographicAttribute$	$\sqcap$	$ProfileAttribute$	
$BehaviourAttribute$	$\sqcap$	$ProfileAttribute$	

La classe **`Segment`** contient plusieurs sous-classes, chaque sous-classe correspondant à un segment marketing correspondant à une campagne publicitaire donnée. Un segment est construit en combinant des éléments "stables" (non-dynamiques) parmi les attributs d'un profil utilisateur. La construction d'un segment avec des règles logiques est discutée dans la Section 3.2.4.2.4. Je donne ci-dessous deux exemples de segments marketing, l'un correspondant à l'exemple de la "maman sportive", l'autre spécifiant le concept de "Geek Android".

$Segment$	$\sqsubseteq$	$\top$	Equation 10
$AndroidGeek$	$\sqsubseteq$	$Segment$	
	$\sqcap$	$hasProfileAttribute.(AndroidUser)$	
	$\sqcap$	$hasTopicalInterest.\{Informatics\_Electronics\}$	
$SportyMom$	$\sqsubseteq$	$Segment$	
	$\sqcap$	$hasProfileAttribute.(Gender\_Woman)$	
	$\sqcap$	$hasProfileAttribute.Family\_Child$	
	$\sqcap$	$hasTopicalInterest.\{Sports\}$	

3.2.4.2.2 Représentation d'informations - Gestion de l'incertitude

Un problème fondamental du profilage est que les conclusions concernant les utilisateurs ne sont pas forcément binaires. Au lieu d'une catégorisation "tranchante", les systèmes de profilage modernes évaluent la probabilité qu'un utilisateur soit intéressé par un sujet ou un produit donné.

Pour gérer ce type d'incertitude, nous avons dû étendre le modèle de base de l'ontologie. Cette extension comprend l'ajout de nouveaux concepts au modèle ontologique, ainsi que d'un formalisme pour la construction de relations quantifiées. Dans le contexte des travaux menés dans cette thèse, le terme "relation quantifiée" fait référence à des relations sémantiques qui sont étendues par en leur associant

une valeur numérique. Cette valeur sera interprétée en tant que pondération et permettra de quantifier la force (ou la faiblesse) de la relation considérée.

Le modèle d'ontologie tel qu'il a été discuté jusqu'à présent correspond à ce qu'on peut appeler une "utilisation normale" des ontologies. Le modèle ontologique présenté modélise les concepts existants dans domaine de la publicité en ligne et définit de manière explicite les relations sémantiques entre ces concepts. Toutefois, pour répondre à l'un des principales questions de recherche poursuivies à travers ce travail de recherche, nous souhaitons que l'ontologie soit utilisée de manière plus dynamique par le système MindMinings et surtout qu'elle permette d'intégrer au processus de profilage les imprécisions (ou incertitudes) associées à l'interprétation de comportements d'utilisateurs sur le Web.

Pour ce faire, nous avons ajouté deux principales extensions au modèle ontologique de base: la classe **`Weight`** et un chaînage entre propriétés.

La classe **`Weight`** a été ajoutée afin de quantifier numériquement l'incertitude. Les instances de cette classe possèdent une propriété de type de données **`hasWeightValue`**, acceptant des nombres à virgules (**`xsd:float`**) et représentant la valeur numérique de l'imprécision.

$Weight$	$\sqsubseteq$	$\top$	Equation 11
	$\sqcap$	$\exists hasWeightValue.(xsd{:}float)$	

La pondération d'une relation entre deux classes de l'ontologie a été implémentée, certes sur la base de la classe Weight, mais surtout à travers trois nouvelles propriétés objet (comme illustré par la Figure 23) **`isRelatedTo-weight`**, **`weight-isRelatedTo`**, **`isRelatedTo`**. Une fois ces propriétés ajoutées, nous avons pu définir le concept de "Relation pondérée" ou "*Weighted Relationship*" (équation 12).

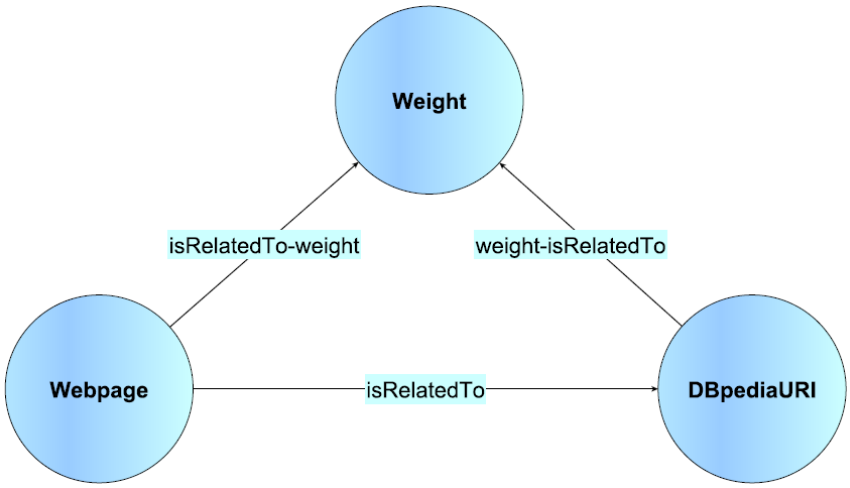


Figure 23. Illustration schématique de la construction d'une relation pondérée entre la Classe Webpage et la classe Universe (Hoppe 2016)

$weightedRelationship$	$\sqsubseteq$	$\top$	Equation 12
$toWeightRelationship$	$\sqsubseteq$	$\top$	
$fromWeightRelationship$	$\sqsubseteq$	$\top$	
$\top$	$\sqsubseteq$	$\forall toWeightRelationship.Weight$	
$\exists fromWeightRelationship.\top$	$\sqsubseteq$	$Weight$	
$isRelatedTo$	$\sqsubseteq$	$weightedRelationship$	
$isRelatedTo\_weight$	$\sqsubseteq$	$toWeightRelationship$	
$\exists\ isRelatedTo\_weight.\top$	$\sqsubseteq$	$Webpage$	
$weight\_isRelatedTo$	$\sqsubseteq$	$fromWeightRelationship$	
$\top$	$\sqsubseteq$	$\forall toWeightRelationship.Webpage$	
$isRelatedTo$	$\equiv$	$isRelatedTo\_weight \circ weight\_isRelatedTo$	

Une fois ce concept ajouté à la TBox et afin de préserver la cohérence de la TBox, nous avons ajoutés les différents triples précisant quelles propriétés (parmi celles déjà présentes dans l'ontologie) deviennent des relations pondérées (voir équation 13).

$\exists weightedRelationship.\top$	$\sqsubseteq$	$\top$	Equation 13
$\top$	$\sqsubseteq$	$\forall weightedRelationship.\top$	
$entailsTopicalInterest$	$\sqsubseteq$	$weightedRelationship$	
$containsKeyword$	$\sqsubseteq$	$weightedRelationship$	
$hasTopicalInterest$	$\sqsubseteq$	$weightedRelationship$	
$isMappedTo$	$\sqsubseteq$	$weightedRelationship$	
$belongsToSegment$	$\sqsubseteq$	$weightedRelationship$	
$hasProfileAttribute$	$\sqsubseteq$	$weightedRelationship$	

3.2.4.2.3 Exploitation du profil – Raisonnement numérique à base de règles logiques

Notre approche pour la construction du profil utilisateur final repose donc sur ces relations pondérées, qui sont utilisées par le système MindMinings pour déduire des caractéristiques haut niveau d'un profil. Toutefois lorsqu'il s'agit de manipuler de telles relations pondérées avec un raisonneur, plusieurs problèmes sont soulevés. Alors qu'interprétant sans problèmes les constructions à la base des langages OWL-DL et SWRL, un raisonneur ne saura pas interpréter nos relations pondérées en tant qu'expressions d'incertitudes. Plus particulièrement, les problèmes de propagation, d'agrégation et de comparaison de valeurs numériques ont été adressés :

- **Propagation de valeurs numériques** - Le système doit être capable de propager des valeurs à travers la structure relationnelle de l'ontologie. En pratique, cette propagation peut prendre deux formes principales:
  - Lorsque l'on doit agréger la valeur d'une relation pondérée avec une relation qui ne l'est pas, cela revient à propager la valeur d'une instance de la classe Weight à une autre instance de la même classe. C'est par exemple le cas pour l'agrégation entre la propriété **`requestsWebpage`** (reliant une instance de la classe **`Hit`** à une instance de la classe **`Webpage`**) et la propriété **`containsRelationToUniverse`** (reliant une instance de la classe **`Webpage`** à une instance de la classe **`Universe`**). Pour simplifier l'agrégation de scores d'intérêts de l'utilisateur, nous avons ajouté une nouvelle relation pondérée, **`entailsTopicalInterest`**. Comme la première partie de cette relation composée n'est pas pondérée, il suffit de propager la pondération associée (valeur de la propriété **`hasWeight`** de l'instance **`Weight`** associée) à la relation **`containsRelationToUniverse`** en tant que valeur de la propriété **`hasWeight`** de l'instance **`Weight`** créée pour la nouvelle propriété.
  - Lorsque la présence d'une valeur numérique doit être interprétée d'une certaine manière. En effet, l'interprétation de telles valeurs numériques n'est pas une chose aisée pour un utilisateur non expert du système ou du domaine. Afin d'assurer une interprétation claire et uniforme des énoncés de l'ontologie, nous avons ajouté deux nouvelles propriétés à la classe **`Universe`**, **`hasStrongInterestIn`** (ajoutée automatiquement pour chaque relation **`hasTopicalInterest`** ayant une pondération supérieure à 0,85) et **`hasWeakInterestIn`** (ajoutée automatiquement pour chaque relation hasTopicalInterest ayant une pondération inférieure à 0,15).
- **Agrégation de valeurs numériques** - lorsqu'il s'agit d'agréger deux relations pondérées, les valeurs numériques associées à chacune d'entre elles doivent être combinées (en appliquant un traitement donné ou une fonction d'agrégation). Deux types de fonctions d'agrégation ont été examinés en particulier, à savoir le comptage et la moyenne. Des fonctions de comptage sont en effet nécessaires pour déterminer le nombre de pages Web consultées par utilisateur ou encore le nombre de centres d'intérêt d'un utilisateur donné. Le calcul d'une moyenne de pondérations s'avère utile lorsqu'il s'agit de quantifier la relation entre un utilisateur et un certain centre d'intérêt, puisque ce type de relation comprend plusieurs relations reliant un utilisateur à une page Web, elle-même reliée à un univers.
- **Comparaison de valeurs numériques** - ceci est nécessaire afin d'évaluer les résultats du processus de profilage et permettre de répondre à des requêtes comme par exemple quels sont les principaux centres d'intérêt d'un utilisateur par ordre de préférence décroissant. Le calcul de valeurs maximales, l'ordonnancement d'éléments ou encore l'instanciation automatique des propriétés

`hasStrongInterest` et **`hasWeakInterest`** sont des fonctionnalités du système nécessitant de comparer des valeurs numériques entre elles.

L'analyse ci-dessus a permis d'identifier les fonctionnalités nécessaires à l'intégration de l'ensemble du processus de profilage dans l'ontologie MindMinings. Ayant choisi le langage SWRL pour la formalisation de nos règles logiques, je présente ci-dessous comment ces fonctionnalités sont implémentées avec SWRL. Une attention particulière est accordée au fait de ne pas dépasser les limites des langages de modélisation utilisés (OWL-DL et SWRL) afin de maintenir l'ontologie décidable.

Pour la propagation de valeurs numériques, notre formalisme requiert deux éléments : a) la valeur à propager, b) la cible de la propagation. Dans le système MindMinings, des règles SWRL sont définies pour chacune des agrégations de propriétés identifiées. Ci-dessous, un exemple de propagation permettant de relier une instance de la classe **`Hit`** directement aux instance de la classe **`Universe`**, présentes dans la page Web concernée par l'instance de **`Hit`**.

```
Hit(?hit), Webpage(?wp), Universe(?uni),
Weight(?w), Weight(?w_res),
hasWeightValue(?w,?v),
requestsWebpage(?hit,?wp),
containsRelationToUniverse-weight(?wp,?w),
weight-containsRelationToUniverse(?w,?uni),
entailsTopicalInterest-weight(?user,?w_res),
weight-entailsTopicalInterest(?w_res,?uni)
   →
hasWeightValue(?w_res,?v)
```

Figure 24. Propagation de valeurs numériques par le biais de règles SWRL (Hoppe 2016)

Pour l'agrégation de valeurs numériques, notre formalisme prend en compte deux types d'agrégation:

a) **L'agrégation de valeurs de relations en série**, c'est-à-dire des relations qui, lorsqu'appliquées de manière séquentielle, permettent de passer d'un concept à un autre. Dans ce cas, la complexité des règles SWRL associées grandit de manière linéaire par rapport au nombre de relations à agréger. Si l'on considère l'agrégation des pondérations associées à n relations, la règle SWRL sous-jacente nécessiterait $6 * n + 4$ formules atomiques (ou atomes) - celles-ci sont listées ci-dessous. Même si la complexité augmente de manière linéaire, ceci a quand même deux impacts sur le système MindMinings: a) plus la base de règles logiques est importante, plus un utilisateur humain aura des difficultés à l'utiliser et à la comprendre, et b) la complexité du processus de raisonnement associé augmente aussi.

- Définition des concepts impliqués dans les n relations (n+1 atomes)
- Définition des instances de la classe **`Weight`** associées aux n relations (n atomes)
- Définition de l'instance de la classe **`Weight`** qui sera la cible de l'agrégation (1 atome)
- Définition de 2 relations partielles, constituant les n relations pondérées considérées (2*n atomes)
- Spécification de la relation **`hasWeightValue`** pour chaque instance de la classe **`Weight`** considérée (n atomes)
- Application itérative de la fonction d'agrégation (n-1 atomes)
- Définition de la relation pondérée cible (2 atomes)
- Définition de la valeur du résultat à associer à l'instance de la classe Weight définie comme cible de l'agrégation (1 atome)

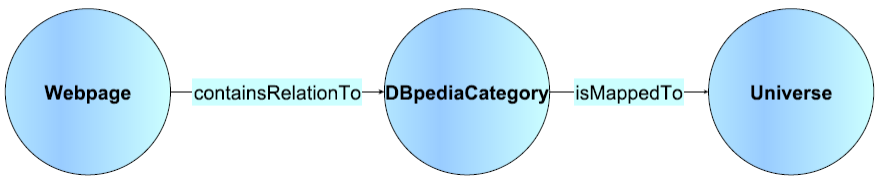


Figure 25. Vue schématique de relations en série – l'agrégation concerne les pondérations associées aux propriétés **`containsRelationTo`** et **`isMappedTo`** (Hoppe 2016)

```
Webpage(?wp),DBpediaCategory(?cat), Universe(?uni),
Weight(?w1),Weight(?w2),Weight(?w),
containsRelationToCategory-weight(?wp,?w1),
weight-containsRelationToCategory(?w1,?cat),
isMappedTo-weight(?cat,?w2),
weight-isMappedTo(?w2,?uni),
hasWeightValue(?w1,?v1),
hasWeightValue(?w2,?v2),
swrlb:multiply(?v3,?v2,?v1)
containsRelationToUniverse-weight(?wp,?w),
weight-containsRelationToUniverse(?w,?uni),
->
hasWeightValue(?w,?v3)
```

Figure 26. Exemple d'agrégation de pondérations de relations en série, en utilisant des règles SWRL (Hoppe 2016)

b) **L'agrégation de valeurs de relations parallèles.** Dans ce cas, la complexité d'une règle SWRL agrégeant n relations parallèles serait de $\frac{n^2}{2} + \frac{5*n}{2} + 5$ et implique les formules atomiques listées ci-dessous. La relation entre nombre de règles à agréger et la régle SWRL associée n'est plus d'ordre linéaire mais devient quadratique. L'impact sur les performances en termes de rapidité de raisonnement sera d'autant plus important.

- Définition de la classe à utiliser en tant que domaine de la relation agrégée (1 atome)
- Définition de la classe à utiliser en tant que portée de la relation agrégée (1 atome)
- Définition des n instances de la classe **`Weight`** à agréger (n atomes)
- Définition de l'instance de la classe **`Weight`** à utiliser pour la relation agrégée (1 atome)
- Spécification des relations **`hasWeightValue`** pour les n instances de la classe **`Weight`** considérées $\sum_1^{n-1} x$
- Application de la fonction d'agrégation, en combinant le n instances de la classe **`Weight`**, sur la base d'une fonction binaire intégrée à SWRL (utilisation de builtins) (n-1 atomes)
- Assignation de la pondération ainsi calculée à l'instance de la classe **`Weight`** définie pour la relation agrégée (1 atome)

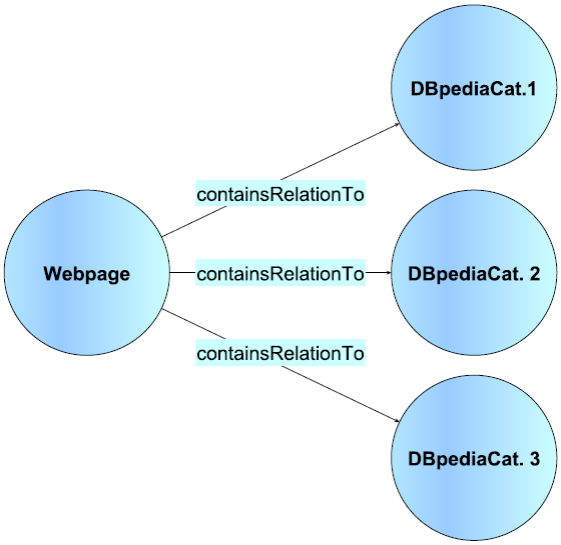


Figure 27. Vue schématique de relations parallèles – l'agrégation concerne les pondérations associées aux différentes relations `containsRelationTo` (Hoppe 2016)

```
Webpage(?wp),DBpediaCategory(?cat),
Weight(?w1),Weight(?w2),Weight(?w),
DifferentFrom(?w1,?w2)
hasWeightValue(?w1,?v1),
hasWeightValue(?w2,?v2),
containsRelationToCategory-weight(?wp,?w1),
weight-containsRelationToCategory(?w1,?cat),
containsRelationToCategory-weight(?wp,?w2),
weight-containsRelationToCategory(?w2,?cat),
containsRelationToCategorySum-weight(?wp,?w),
weight-containsRelationToCategorySum(?w,?uni),
swrlb:add(?v3,?v2,?v1)
->
hasWeightValue(?w,?v3)
```

Figure 28. Exemple d'agrégation de pondérations de relations en parallèle, en utilisant des règles SWRL (Hoppe 2016)

Enfin, pour la comparaison de valeurs numériques, nous pouvons utiliser les fonctions intégrées à SWRL (*SWRL builtins*). Un exemple de telle règle SWRL est donné ci-dessous (voir Figure 29). Elle permet d'instancier une nouvelle relation entre une instance de la classe **`Webpage`** et une instance de la classe **`DBpediaCategory`** représentant le domaine dominant de la page Web.

```
Webpage(?wp),DBpediaCategory(?cat1), DBpediaCategory(?cat2),
Weight(?w1),Weight(?w2),Weight(?w),
containsRelationToCategory-weight(?wp,?w1),
weight-containsRelationToCategory(?w1,?cat1),
containsRelationToCategory-weight(?wp,?w2),
weight-containsRelationToCategory(?w2,?cat2),
swrlb:notEqual(?cat1,?cat2),
hasWeightValue(?w1,?v1),
hasWeightValue(?w2,?v2),
swrlb:lessThan(?v2,?v1)
->
containsMainCategory(?wp,?cat1)
```

Figure 29. Règle SWRL permettant de calculer une valeur maximale (utilisation de **`swrlb:notEqual`** et **`swrlb:lessThan`**) (Hoppe 2016)

Pour conclure cette section, la plupart des éléments nécessaires ont été modélisés par le biais de règles SWRL, à une seule exception. En effet, SWRL ne permet pas aux règles d'inclure un raisonnement non-monotone. SWRL adopte l'hypothèse du monde ouvert et de ce fait, certaines fonctionnalités sont impossibles à implémenter dans ce contexte. C'est notamment le cas lorsque l'on souhaite implémenter des règles de comptage d'éléments. Une alternative à cette limite serait d'utiliser l'extension de SWRL appelée SQWRL (*Semantic Query-Enhanced Web Rule Language*) pour modéliser des règles de comptage d'éléments. SQWRL possède un opérateur sqwrl:count, qui prend un seul argument, à savoir une instance d'une classe. Plusieurs exemples illustrant comment il est possible de compter des instances avec cet opérateur sont disponibles à partir de la documentation officielle de SQWRL: https://github.com/protegeproject/swrlapi/wiki/SQWRLCore#Counting

3.2.4.2.4 Mise à jour du profil - Construction de segments

La mise à jour d'un profil utilisateur revient à modifier des segments marketing existants ou encore à en définir des nouveaux. . Pour modéliser de tels segments, de commun accord avec les experts du domaine, nous avons choisi d'utiliser encore le langage SWRL. Les segments marketing peuvent impliquer ou non des relations pondérées (un exemple est fourni pour chaque cas de figure ci-dessous).

```
BID(?user), Family_Child(?child),
hasTopicalInterest(?user, U-sports),
hasProfileAttribute(?user, ?child),
hasProfileAttribute(?user, Gender_Woman)
 ->
SportyMom(?user)
```

Figure 30. Règle SWRL définissant le segment "Maman sportive" (Hoppe 2016)

```
BID(?user), Weight(?weight),
hasTopicalInterest-weight(?user,?weight),
weight-hasTopicalInterest(?weight,U-Sports),
hasWeightValue(?weight,?value),
swrlb:greaterThan(?value,0.85f)
->
SportsFanatic(?user)
```

Figure 31. Règle SWRL définissant le segment "Fan de sports" (Hoppe 2016)

Les règles SWRL de haut niveau (décrites ci-dessus) illustrent comment est implémentée la segmentation des utilisateurs, c'est-à-dire l'association entre instances d'utilisateurs individuels à des segments de clientèle (identifiant les utilisateurs les plus à mêmes d'être intéressés par une publicité donnée). Ces segments marketing correspondent à l'étape finale du processus de profilage, à savoir sa mise à jour dynamique, à travers des règles logiques. L'adaptation de la base de règles permet d'adapter le système aux nouveaux paramètres de l'application lorsque cela est nécessaire (pour chaque nouvelle campagne publicitaire). Une mise à jour dynamique contraste avec l'approche utilisée par l'entreprise au moment de notre collaboration. En effet, ayant fait le choix d'une approche à base d'apprentissage supervisé, l'ajustement du modèle de profil à chaque nouvelle campagne publicitaire était réalisé manuellement par les experts du domaine.

3.2.4.3 Prototype

3.2.4.3.1 Présentation

Étant donné la complexité des tâches à réaliser, ainsi que le nombre important d'informations et données à traiter, une approche semi-automatique a été envisagée. A travers le prototype présenté ci-dessous, nous avons implémenté une preuve de concept d'un nouvel écosystème pour la publicité en ligne, reposant sur l'exploration de nouvelles techniques pour l'analyse de contenus ainsi que pour la modélisation des connaissances. Ce prototype permet deux types de personnalisation:

- **Personnalisation du profil utilisateur** - l'ontologie de domaine conçue n'intègre que des éléments pertinents pour le profilage d'utilisateurs
- **Personnalisation des interactions** entre les différents modules du système - afin de les adapter à la procédure contractuelle suivie par le partenaire industriel, mais aussi pour améliorer le traitement temps-réel des données

Afin de mieux comprendre les choix faits pour le développement de ce prototype, il est nécessaire de préciser le contexte dans lequel il devait être utilisé par l'entreprise partenaire. En effet, cette dernière offre des services à l'ensemble des acteurs de l'écosystème publicitaire. Toutefois, le système MindMinings ne concerne qu'un sous-domaine de leur expertise: la modélisation du profil des utilisateurs en fonction de leur historique de navigation Web. Le système MindMinings (et le service de profilage associé) ne s'adresse qu'aux éditeurs en ligne ayant souscrit à un contrat avec l'entreprise partenaire.

Les différents journaux de navigation Web constituent le point de départ du processus de profilage: le système lit les champs de ces journaux, stocke les informations directement exploitables et procède à l'analyse des contenus les plus complexes. Ces éléments complexes sont avant tout les URLs des contenus Web consultés par l'utilisateur. Ils constituent la principale source d'information et font l'objet d'analyses souvent longues. La nécessité de résultats rapides conduit souvent à une limitation de la profondeur de ces analyses.

La manière dont interagissent l'éditeur et l'entreprise partenaire permet une adaptation de l'analyse de ces journaux. En effet, l'éditeur n'est pas en mesure de suivre l'activité de navigation complète des utilisateurs. L'historique côté serveur ne capture que les éléments qui ont été demandés à partir des serveurs de l'éditeur. Par conséquent, l'ensemble d'URLs apparaissent dans les journaux est également limité au contenu de l'éditeur. De ce fait, cela nous a permis d'implémenter une analyse asynchrone des différentes pages Web identifiées par leurs URLs respectives. Ainsi, au lieu d'analyser les contenus à la volée lors de la réception des journaux de navigation, il est possible de les prétraiter. Dès qu'un contrat est signé avec un éditeur, les sites Web des nouveaux clients sont parcourus. Les pages Web trouvées sont analysées et leurs informations de contenu sont stockées dans la base de connaissances.

Ensuite, lorsque les pages Web réellement parcourues par l'utilisateur sont enregistrées dans le journal de navigation, il est uniquement nécessaire de connecter l'instance utilisateur correspondante dans la base de connaissances avec les informations de contenu appropriées. Cette adaptation, couplée à la base sophistiquée de règles logiques et un raisonneur, permet de déduire automatiquement le profil utilisateur.

L'architecture du prototype MindMinings est illustrée par la figure ci-dessous. Les modules illustrés avec un fond de couleur représentent des tâches réalisées de manière asynchrone (analyse des pages Web

et gestion de la base de règles), alors que les modules illustrés avec un fond blanc représentent des tâches réalisées de manière synchrone.

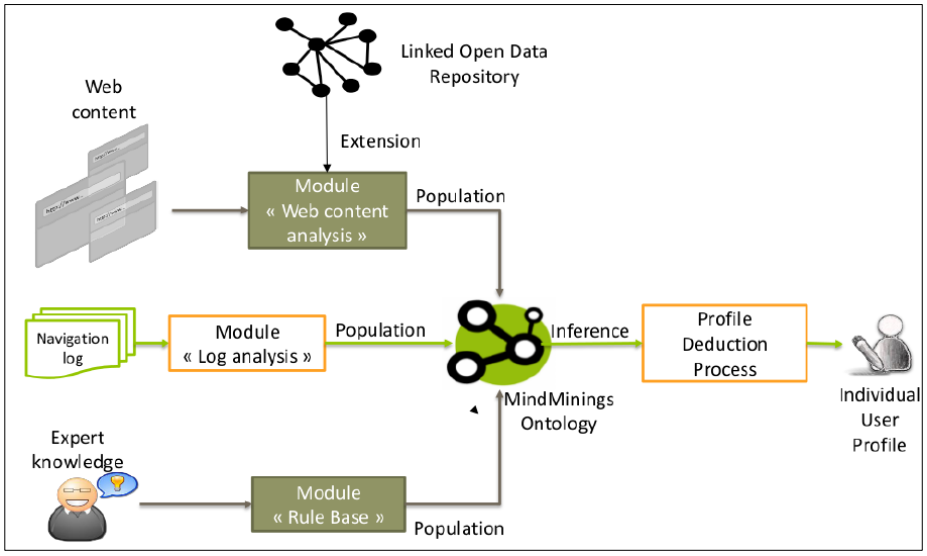


Figure 32. Architecture du système MindMinings (Hoppe 2016)

La Figure 33 illustre comment les concepts de l'ontologie intervenant dans les modules et sous-systèmes illustrés ci-dessus (voir Figure 32).

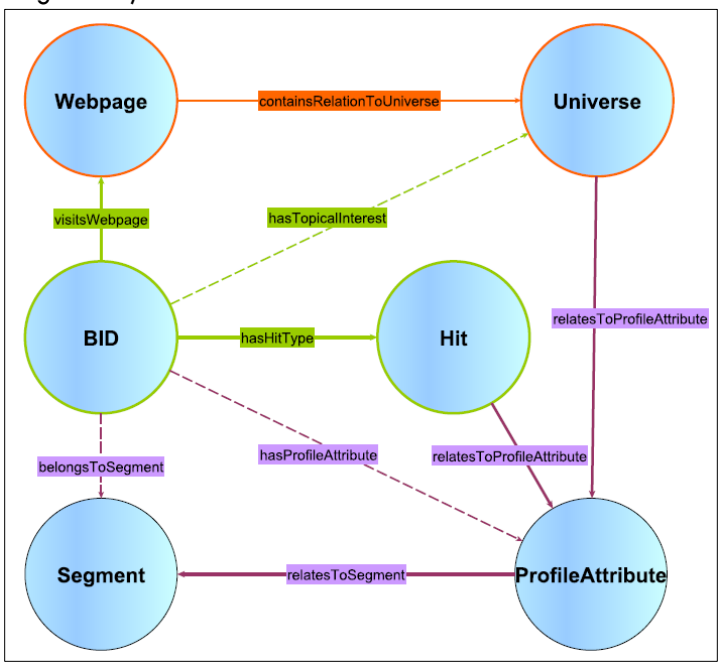


Figure 33. Classes impliquées dans la construction du profil – orange pour les éléments du module "Web content analysis", vert pour le module "Log analysis" et violet pour les informations déduites à partir des règles du module "Expert rule base". Les lignes continues indiquent des relations qui sont ajoutées par les modules respectifs, alors que les lignes en pointillés précisent les relations inférées. (Hoppe 2016)

Les tâches des différents modules sont listées dans le Tableau 13.

Module	Description
Analyse de contenus Web "Web content analysis"	L'analyse asynchrone des contenus Web permet d'utiliser des étapes d'analyse plus complexes. Dans la version prototype, le module intègre une extraction statistique des mots-clés et un cadre pour la reconnaissance des entités nommées. Les informations récupérées sont ensuite améliorées à l'aide de ressources externes provenant du cloud Linked Open Data. Ce profil de page étendu est

	ensuite utilisé pour classer le contenu dans l'ensemble des rubriques de domaine personnalisée
Analyse de journaux de navigation "Log analysis"	Parcourt le journal de navigation et transmet les informations extraites à l'ontologie
Base de règles "Rule base"	Réunit les informations obtenues à partir de l'analyse de pages Web et du journal de navigation pour construire le profil utilisateur.

Tableau 13. Tâches assurées par les différents modules illustrés à la Figure 32

3.2.4.3.2 Interfaces

Un des derniers objectifs en termes de développement fut la conception d'interfaces intuitives pour que les experts métiers puissent spécifier leurs connaissances et composer de nouveaux segments marketing. Pour ce faire, trois interfaces ont été développées:

> Pour la **gestion de règles** - L'interface (Figure 34) se limite aux règles spécifiques à une campagne donnée ainsi qu'à certains concepts spécifiques (e.g. les différentes instances des sous-classes de la classe Hit). Lorsqu'une règle est sélectionnée dans la liste, ses informations sont affichées sur le côté droit de l'écran (e.g. nom de la règle, son annotation, son corps et sa tête). Les boutons dans la partie inférieure de l'écran permettent de modifier la règle sélectionnée ou de l'effacer de la base de connaissances.

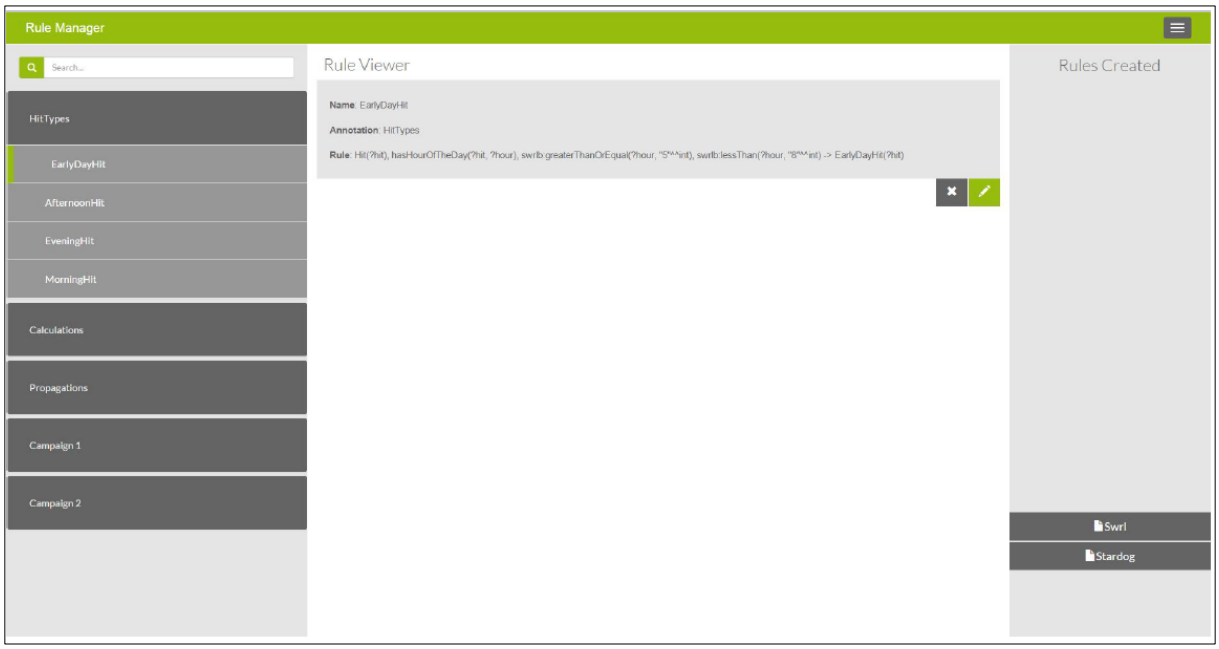


Figure 34. Interface pour la gestion des règles SWRL (Hoppe 2016)

> Pour la construction de règles - Lorsque dans l'interface précédente on choisit de modifier une règle existante, cela ouvre la fenêtre d'édition de règles logiques illustrée par la Figure 35. Sur le côté gauche on retrouve les différents atomes pouvant être utilisés pour la composition de règles, à savoir les classes et les propriétés de l'ontologie, mais aussi l'ensemble des constructions SWRL (builtins) tels que **`swrlb:add`** ou **`swrlb:divide`**. L'utilisateur compose sa règle en glissant et en déposant ces atomes depuis la gauche vers le centre de l'écran. Les zones "If" et "Then" représentent l'en-tête, et, respectivement, le corps de la règle. C'est ce type d'interface qui doit être utilisé pour spécifier les nouveaux segments marketing ou en modifier des anciens.
> Pour l'exécution confinée de règles - Le système MindMinings permet de tester les implications logiques produites par une nouvelle règle, et ce avant de l'ajouter à la base de connaissances. Pour ce faire, côté serveur, une copie de la base de connaissances est produite, et c'est à cette copie que la nouvelle règle est ajoutée. L'exécution de la règle logique peut dès lors être testée sur la copie de la base de connaissances, sans risquer un effet de bord non prévu. Si l'exécution n'introduit aucun

problème, alors la règle peut être ajoutée à la base de connaissances d'origine. La Figure 36 illustre un exemple d'une règle ayant produit une erreur, signalée par le raisonneur Pellet utilisé (mauvaise utilisation de **swrlb:notEqual**). Pour des raisons de sécurité, seules les règles logiques dont l'exécution a été vérifiée peuvent être ajoutées à la base de connaissances.

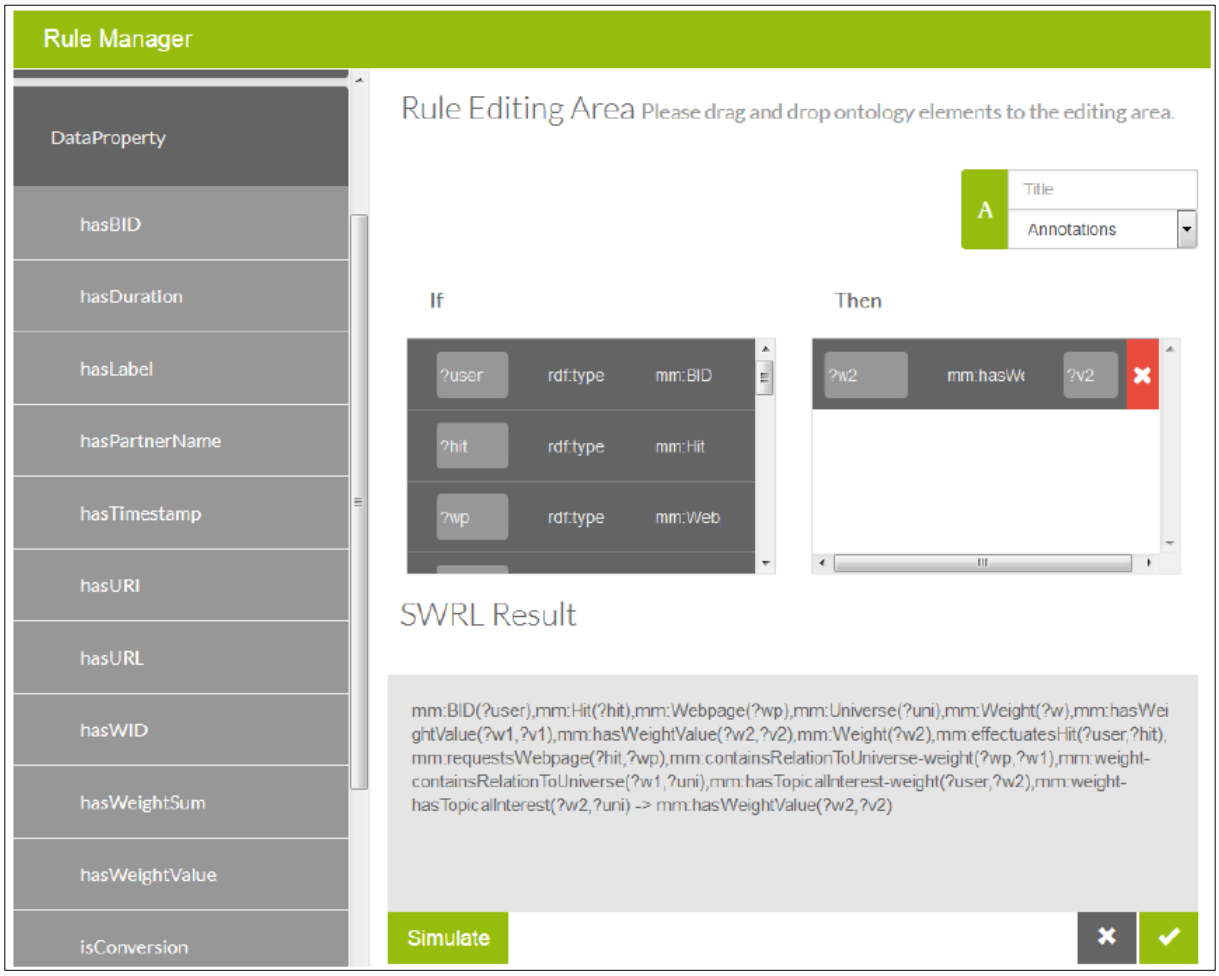


Figure 35. Interface pour la construction de règles SWRL par glisser-déposer (Hoppe 2016)

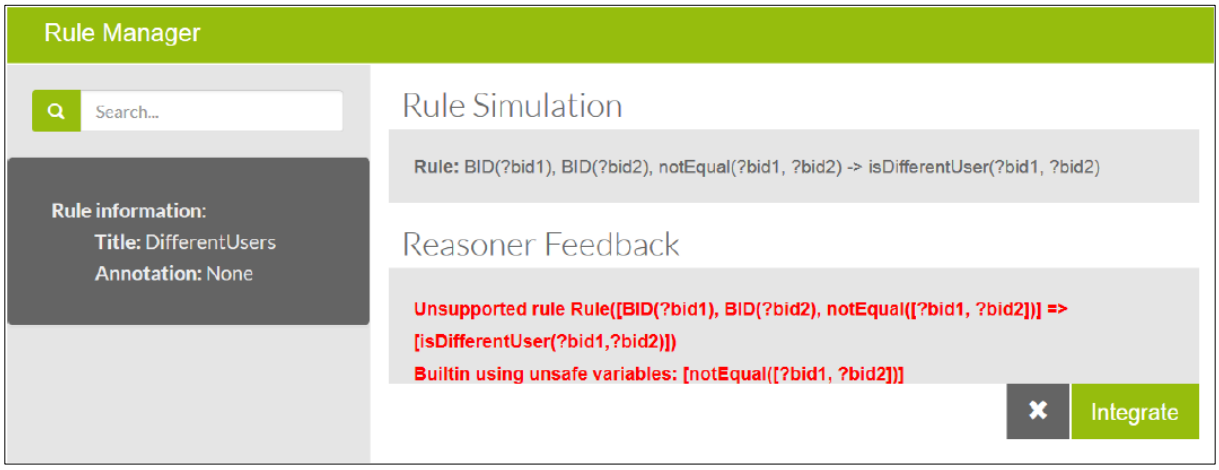


Figure 36. Exécution confinée de règles SWRL (Hoppe 2016)

3.2.4.3.3 Évaluation

3.2.4.3.3.1 Classification pages Web

La performance du processus de classification des pages Web a été testée au travers de trois approches:

- Classification simple à base de mots-clés (S) - Utilise uniquement le jeu de mots-clés initial défini pour une catégorie donnée. L'ensemble de mots-clés associé à chaque catégorie est comparé à l'ensemble des mots-clés extraits de la page Web. A chaque fois qu'un appariement est trouvé, le score de la catégorie considérée augmente. A la fin de ce processus, la page Web est classée comme appartenant à la catégorie ayant rencontré le plus grand score. L'ensemble des scores est ensuite normalisé à un.
- Classification étendue à base de mots-clés (E) - Utilise le jeu de mots-clés étendu avec des relations sémantiques vers des ressources de DBpedia. Le processus de classification en lui-même suit les mêmes étapes que précédemment: (i) les mots-clés de la catégorie sont comparés aux mots-clés de la page Web; (ii) Les notes de la catégorie sont augmentées du score de chaque mot-clé pour lequel une correspondance est trouvée; (iii) Les notes finales des catégories sont normalisées à un.
- Classification MindMinings (à base d'ontologie) (O) - Les différentes catégories et les mots-clés extraits de la page Web sont identifiés par leurs URIs DBpedia, et l'ensemble de ces instances sont ajoutées à la base de connaissances MindMinings. Le résultat de la classification exploite la structure relationnelle de l'ontologie. Une comparaison est effectuée entre les URIs de haut niveau récupérées pour les mots-clés de l'article et les URIs calculées en fonction du jeu de mots-clés initial. Chaque correspondance augmente le score de l'univers considéré. En plus du score des mots-clés, cette méthode prend également en compte la pondération associée à la relation entre l'URI DBpedia et l'univers considéré. Le score de l'univers est augmenté du produit de ces deux valeurs. Au final, les valeurs de score pour tous les univers sont normalisées à un.

Ces trois classificateurs ont été appliqués sur un ensemble de 28 pages Web extraites à partir de sites d'actualités en français. N'ayant pas trouvé de référence avec laquelle comparer nos résultats, l'ensemble des 34 pages Web ont été classées à la main, par rapport à l'ensemble des univers considérés dans le prototype MindMinings. Cela a permis d'associer entre 4 et 5 univers par page Web, avec des scores variant de 0,1 (rapport le plus faible) à 0,6 (forte prévalence de l'univers thématique respectif).

L'analyse générale des résultats de classification a révélé que le classificateur à base d'ontologie fournit une gamme de classifications de catégories plus large que les deux alternatives basées sur des mots-clés. Les classificateurs par mots clés trouvaient généralement deux ou trois univers avec de très hauts scores, alors que les scores des autres univers restaient à 0. Contrairement à cela, le classificateur basé sur l'ontologie est plus à même de répartir les scores sur au moins cinq catégories d'univers.

Article no	RMSE S	RMSE C	RMSE O
0	0,02975	0,02975	0,0077
1	0,00572	0,00937	0,00566
2	0,01568	0,04593	0,00631
3	0,01106	0,00668	0,00346
4	0,02577	0,02577	0,02573
5	0,00884	0,00884	0,00615
6	0,04077	0,04077	0,01579
7	0,00746	0,00846	0,0007
8	0,00631	0,00631	0,01659
9	0,03028	0,03028	0,00344
10	0,00446	0,00446	0,02283
11	0,02499	0,00766	0,01302
12	0,04959	0,00749	0,00725
13	0,01538	0,01538	0,01934
14	0,01541	0,01541	0,00625
15	0,01892	0,01245	0,02371
16	0,01841	0,01988	0,00885
17	0,04692	0,04692	0,03676
18	0,00505	0,00763	0,00358
19	0,00669	0,00669	0,00431
20	0,00672	0,00672	0,0073
21	0,00604	0,00667	0,00107
22	0,02904	0,02904	0,02725
23	0,03333	0,03333	0,01149
24	0,00384	0,00384	0,00118
25	0,01233	0,01233	0,00356
26	0,02191	0,0138	0,01514
27	0,00452	0,00452	0,00703
28	0,01923077	0,01923077	0,03821201

Tableau 14. Evaluation des trois approches de classification (Hoppe 2016)

Le Tableau 14 contient les résultats comparatifs entre les trois approches de classification testées. La qualité de la classification produite est mesurée en tant que *Root Mean Squared Error* (RMSE), en respectant l'annotation manuelle ayant servi de référence. En ce sens, une valeur plus basse dénote une meilleure performance. Les entrées en rouge correspondent aux pires résultats de classification pour l'article considéré. Cependant, dans de nombreux cas, la diversification introduite par le classificateur basé sur l'ontologie semble conduire à de meilleurs résultats, comme le suggère la somme des erreurs, présentée dans le Tableau 15.

$\sum RMSE_S$	$\sum RMSE_E$	$\sum RMSE_O$
0,62364	0,58482	0,41994

Tableau 15. Somme des erreurs pour chaque approche de classification testée.

Pour tester cette hypothèse, le ré échantillonnage non paramétrique ou bootstrap[16] (Efron 1979) est appliqué aux taux d'erreur expérimentaux de tous les classificateurs. Les valeurs expérimentales sont ré-échantillonnées à n = 500 et utilisées pour déterminer les intervalles de confiance pour la moyenne et la médiane des échantillons de données. Le Tableau 16 affiche les résultats pour les valeurs de confiance 0,95, 0,9 et 0,8.

[16] Cette appellation est inspirée du baron de Münchhausen (Rudolph Erich Raspe) qui se sortit de sables mouvants par traction sur ses tirants de bottes. En France "bootstrap" est parfois traduit par à la Cyrano (acte III, scène 13) en référence à ce héros qui prévoyait d'atteindre la lune en se plaçant sur une plaque de fer et en itérant le jet d'un aimant.

Coeff. confiance	Résultats	Classification S	Classification C	Classification O
0,95	Moyenne a	$0{,}044 < a < 0{,}0572$	$0{,}0441 < a < 0{,}0585$	$0{,}0365 < a < 0{,}0489$
	Médiane m	$0{,}0434 < m < 0{,}0572$	$0{,}0409 < m < 0{,}0577$	$0{,}0337 < m < 0{,}0481$
0,9	Moyenne a	$0{,}0451 < a < 0{,}0563$	$0{,}0450 < a < 0{,}0577$	$0{,}0370 < a < 0{,}0475$
	Médiane m	$0{,}0337 < m < 0{,}0481$	$0{,}0443 < m < 0{,}0547$	$0{,}0462 < m < 0{,}0615$
0,8	Moyenne a	$0{,}0462 < a < 0{,}0551$	$0{,}0464 < a < 0{,}0562$	$0{,}0379 < a < 0{,}0461$
	Médiane m	$0{,}0365 < m < 0{,}0447$	$0{,}0430 < m < 0{,}0539$	$0{,}0481 < m < 0{,}0577$

Tableau 16.Résultats d'évaluation : valeurs moyenne et médiane pour les différents classifieurs utilisés, en utilisant un rééchantillonage à n=500 ainsi que différents coefficients de confiance.

Pour rendre les résultats plus compréhensibles, la Figure 37 et la Figure 38 montrent les intervalles de confiance pour les différents classificateurs, pour les valeurs de confiance de 0,95 et respectivement 0,8. Il apparaît clairement que la différence entre les erreurs des classificateurs n'est pas statistiquement significative pour une confiance de 0,95, car les intervalles de confiance se chevauchent. Cependant, en réduisant la confiance demandée à une valeur de 0,8, les intervalles se détachent.

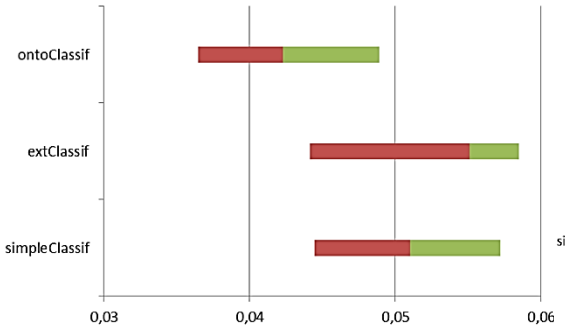


Figure 37. Résultats de l'évaluation des différents classifieurs avec un coefficient de confiance à 0,95

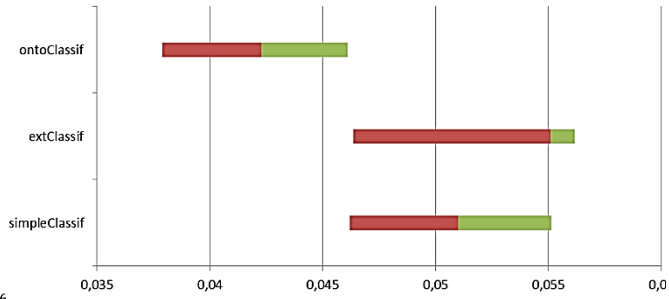


Figure 38. Résultats de l'évaluation des différents classifieurs avec un coefficient de confiance à 0,8

Sur un plan qualitatif, l'application de l'approche de classification basée sur l'ontologie conduit à une diversification des résultats. Au lieu de quelques affectations de catégories très ciblées, le classificateur basé sur l'ontologie produit cinq à six affectations, chacune avec un score inférieur. En conséquence, le classificateur donne de meilleurs résultats sur des contenus qui présentent une diversité de sujets ou contiennent un vocabulaire plutôt non spécifique. Ce qui correspond tout à fait au monde du Web.

Les erreurs de classification ont été examinées davantage en utilisant une méthode de ré-échantillonnage (*bootstrapping*). Les intervalles de confiance résultants suggèrent que, sur un niveau de confiance à 95%, l'amélioration obtenue par le classificateur basé sur l'ontologie n'est pas statistiquement pertinente. Ces résultats peuvent être améliorés si on prend en compte l'ensemble des choix de conception influant sur les résultats de la classification. C'est notamment le cas du processus d'extension des univers (et des mots-clés en général) avec les URIs correspondantes sur DBpedia. Jusqu'à présent, notre algorithme utilise un ensemble de propriétés collectées manuellement pour la récupération des URI de mots-clés et des extensions d'URI. Tester davantage de propriétés pourrait améliorer la qualité des URI dans la description de l'univers et, par la suite, les performances du classificateur. De plus, la redondance sémantique au niveau de la structure même de DBpedia peut poser problème. Les différents prédicats utilisés ont peut-être des équivalents qui sont davantage utilisés pour spécifier des relations au sein de DBpedia. Il s'agirait de spécifier des relations d'équivalence entre ces propriétés, et pouvoir exploiter ces équivalences lors du processus d'extension.

3.2.4.3.3.2 Profilage utilisateurs

Pour la validation de notre approche de profilage utilisateur, nous nous sommes concentrés sur deux principales questions:

> L'approche est-elle fonctionnelle, c'est-à-dire qu'elle fournit des résultats corrects basés sur des valeurs d'entrée connues?
> L'approche passe-t-elle l'échelle, c'est-à-dire peut-elle fournir des résultats corrects même lorsqu'elle est confrontée à de grandes quantités de données?

Produire un profil utilisateur comprend des processus de complexité variable - à partir de la déduction des sous-types corrects du concept `Hit`, à la dérivation des pondérations des relations liant un utilisateur à des univers, des attributs de profil ou encore des segments marketing. Pour plus de clarté et de compréhensibilité, notre évaluation se concentre sur une sous-dérivation de cet ensemble de processus.

En effet, la déduction des pondérations sur les relations entre un utilisateur individuel et des instances d'univers (ou de centres d'intérêt) représente la déduction la plus complexe implémentée dans le prototype MindMinings. Les règles SWRL sous-jacentes incluent, en fonction du nombre d'instances de données considérées, plusieurs comparaisons d'entités, conditions sur des relations et opérations arithmétiques. Le calcul de la pondération finale implique des instances de 4 classes différentes, ainsi que les relations entre elles (voir Figure 39).

En conséquence, ce processus de déduction a été choisi pour tester la robustesse du système de profilage.

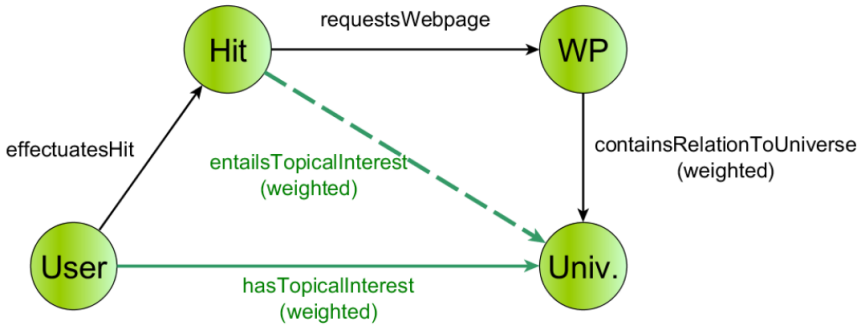


Figure 39. Concepts impliqués dans la déduction d'un profil utilisateur.

Malheureusement les différents tests n'ont pas pu être réalisés dans des conditions industrielles, uniquement en laboratoire. L'ensemble des tests a été effectué sur une seule machine, un ordinateur portable Dell Latitude E5530, avec un processeur Intel(R) i5-3320M, avec une fréquence processeur à 2.60GHz, 8 GB de mémoire RAM et utilisant le système d'opération Windows 7 Professional (64bit), Service Pack 1. Pour la manipulation de l'ontologie, nous avons utilisé l'API OWL à la fois pour le peuplement de l'ontologie et pour la génération de règles SWRL complexes. L'éditeur Protégé (version 4.3.0 build 304) en combinaison avec le raisonneur Pellet 2.2.0 a été utilisé pour afficher les résultats du raisonnement.

Nous avons testé premièrement les résultats de profilage d'un point de vue qualitatif. Pour ce faire, nous avons ajouté à la base de connaissances trois exemples d'utilisateurs, avec un historique de pages Web (entre 2 et 3 pages vues par utilisateur) et déjà catégorisées (pour simplifier, uniquement 5 catégories ont été utilisées). Le Tableau 17 illustre les éléments ajoutés à la base de connaissances. On a ensuite exécuté le processus de déduction des pondérations sur les relations entre chaque utilisateur et les différents centres d'intérêt (univers) considérés.

Utilisateurs	Pages Web	Classifications
Utilisateur 1	pw_{11}	Health = 0,3, Society = 0,4, Sports = 0,2, Work = 0,1
	pw_{12}	Society = 0,2 ; Sports = 0,4 ; Work = 0,4
	pw_{13}	Health = 0,5 ; Society = 0,4 ; Sports = 0,1
Utilisateur 2	pw_{21}	Health = 0,8 ; Sports = 0,2
	pw_{22}	Health = 0,3 ; Society = 0,3 ; Travel = 0,4
	pw_{23}	Health = 0,3 ; Work = 0,7
Utilisateur 3	pw_{31}	Health = 0,5 ; Sports = 0,3 ; Work = 0,2
	pw_{32}	Health = 0,3 ; Work = 0,7

Tableau 17. Evaluation qualitative du processus de profilage – valeurs d'entrée

Le Tableau 18 indique les résultats pour chaque univers. La première valeur de chaque cellule donne la valeur générée par le raisonneur de l'ontologie, la seconde cellule contient le résultat attendu. On peut voir qu'à l'exception de certaines erreurs d'arrondi, les résultats réels correspondent aux valeurs attendues.

Univers	Résultats	User1	User2	User3
Health	Valeur obtenue	0,2666	0,4666	0,4
	Valeur attendue	0,2667	0,4667	0,4
Society	Valeur obtenue	0,3333	0,1	0
	Valeur attendue	0,3333	0,1	0
Sports	Valeur obtenue	0,2333	0,0666	0,15
	Valeur attendue	0,2333	0,0667	0,15
Travel	Valeur obtenue	0	0,1333	0
	Valeur attendue	0	0,1333	0
Work	Valeur obtenue	0,1666	0,2333	0,45
	Valeur attendue	0,1667	0,2333	0,45

Tableau 18. Evaluation qualitative du processus de profilage – résultats pour les trois utilisateurs test

S'en est suivi une évaluation quantitative des performances du prototype. Bien évidemment, n'ayant pas bénéficié d'un environnement industriel avec des infrastructures robustes et des machines puissantes, notre implémentation a impacté d'une manière négative ces résultats. Nous avons toutefois identifié trois paramètres pouvant influer les performances du prototype e.g. le nombre d'univers considéré, le nombre d'instances de la classe Hit, le nombre d'utilisateurs. Afin d'obtenir des résultats plus complets, nous avons réalisé trois séries de tests, et pour chaque série, seul un paramètre sur les trois avait une valeur variable, les autres ayant des valeurs fixes. Cela nous a permis d'obtenir les résultats illustrés par le Tableau 19.

			Temps de classification (en ms)						
#Users	**#Univers**	**#PagesWeb**	Itération 1	Itération 2	Itération 3	Itération 4	Itération 5	**Moyenne**	**Écart-type**
1	1	3	*56*	*70*	*68*	*57*	*63*	62,8	5,04
1	2	3	*205*	*158*	*138*	*146*	*137*	156,8	19,76
1	3	3	*161*	*154*	*144*	*131*	*166*	151,2	10,96
1	4	3	*243*	*334*	*466*	*408*	*432*	376,6	70,48
1	5	3	*1256*	*1216*	*1185*	*1210*	*1302*	1233,8	36,16
1	6	3	*2360*	*1868*	*1839*	*1794*	*1655*	1903,2	182,72
1	7	3	*5861*	*4607*	*4259*	*4189*	*5398*	4862,8	613,36
#Users	**#Univers**	**#PagesWeb**	Itération 1	Itération 2	Itération 3	Itération 4	Itération 5	**Moyenne**	**Écart-type**
1	5	3	*56*	-	-	-	-	56	0
2	5	3	*1944*	*2652*	*2328*	*3055*	*1658*	2327,4	421,12
3	5	3	*5989*	*5362*	*5395*	*5377*	*6050*	5634,6	307,92
4	5	3	*15477*	*15571*	*17600*	*15688*	*13986*	15664,4	783,68
#Users	**#Univers**	**#PagesWeb**	Itération 1	Itération 2	Itération 3	Itération 4	Itération 5	**Moyenne**	**Écart-type**
1	1	2	*56*	*142*	*70*	*68*	*57*	78,6	25,36
1	1	3	*53*	*54*	*36*	*44*	*41*	45,6	6,32
1	1	4	*247*	*168*	*187*	*208*	*139*	189,8	30,16
1	1	5	*4791*	*5018*	*3338*	*3585*	*3209*	3988,2	733,04

Tableau 19. Evaluation quantitative du processus de classification.

Les résultats montrent que c'est le nombre d'utilisateurs qui influe le plus sur le temps nécessaire à la classification. Les résultats obtenus sont en grande partie justifiés par l'environnement de test utilisé, à savoir OWL API et Protégé, ce choix nous ayant été dicté par le fait que les magasins de triples

disponibles à l'époque ne prenaient pas en charge le raisonnement du fait de la complexité du modèle utilisé. Des approches comme le sélecteur de règles logiques (développé dans le cadre de la thèse de M. MENDES de FARIAS et présenté à la section 0) auraient pu aider à améliorer les temps nécessaires à la classification. De plus, les nouvelles versions de magasins de triples, e.g. Stardog, exploitent de plus en plus la puissance de l'informatique en nuage pour accélérer les différents traitements et processus de raisonnement.

3.2.5 Rayonnement scientifique

Au total, le rayonnement scientifique associé aux recherches présentées dans ce chapitre a donné lieu à:

- 2 articles publiés dans des Revues Internationales Sélectives à Comité de Lecture [RIS6], [RIS7]
- 4 communications à des congrès internationaux à comité de sélection et actes publiés [CI14], [CI16], [CI19], [CI20]
- 1 chapitre d'ouvrage [CH6]

A côté de ces publications, plusieurs rapports de recherche internes au projet (et fournis à l'entreprise partenaire) ont aussi été rédigés. Je donne ci-dessous une liste non exhaustive des rapports produits:

- Étude sur les standards existants pour la classification de concepts e.g. classifications NAICS, UNSPSC, GoodRelations, eClass, DMoz, Yahoo!, eOTD - octobre 2012 - 9 pages
- Etat de l'art sur les approches existantes permettant la classification de pages Web notamment les méthodes existantes permettant d'extraire des sous-parties structurées d'une page Web et les différents algorithmes permettant de les traiter (e.g. TF/IDF, Latent Semantic Indexing) - janvier 2013 - 18 pages
- Etat de l'art sur les technologies du Web sémantique - janvier 2013 - 16 pages
- Étude sur le langage de règles logiques SWRL - juillet 2013 - 8 pages
- Etat de l'art sur la logique floue (e.g. opérations floues, inférence en logique floue) et les approches permettant d'intégrer des incertitudes dans des ontologies - février 2013 - 8 pages
- Etat de l'art sur le profilage utilisateur à base d'ontologies - avril 2013 - 12 pages
- Etat de l'art sur les mesures de similarité et les distances sémantiques - juin 2013 - 35 pages
- Rapport sur l'apprentissage supervisé - novembre 2013 - 9 pages
- Etat de l'art sur la lemmatisation (fonctionnalités, plateformes TreeTagger, CST Lemmatizer, Freeling, Flemm, etc) - novembre 2013 - 17 pages

Ces rapports ont été transmis à l'entreprise partenaire et ont fait partie des transferts de compétences réalisés. Plusieurs réunions en présentiel ont permis de former les experts de l'entreprise aux technologies du Web sémantique, ainsi que de répondre à leurs interrogations formulées lors des réunions projet hebdomadaires. Cela a permis de mettre en place une formation à destination des industriels traitant des technologies du Web sémantique, des données liées et notamment des capacités de raisonnement permises par les différents langages d'ontologies. J'ai dispensé cette formation à des industriels (e.g. Groupe SEB, CSTB).

3.2.6 Conclusions et ouvertures

3.2.6.1 Conclusions

Cette partie a résumé les principaux travaux de recherche menés dans le domaine du profilage d'utilisateurs sur la base de leur historique de navigation Web. Ces travaux rejoignent le premier axe des orientations de recherche, à savoir "l'utilisateur" et peuvent être considérés en tant qu'élément de réponse au verrou scientifique traitant de l'impact de la modélisation sur l'implémentation. La principale contribution des recherches présentées dans ce chapitre est de permettre d'appréhender de manière homogène l'ensemble des composants intervenant dans un système de profilage utilisateur. L'intérêt de l'approche est qu'à travers la notion de relation pondérée, elle permet d'exprimer un degré d'incertitude par rapport à un énoncé donné (triple). De plus, elle fournit un mécanisme pouvant classifier les pages du Web selon une taxonomie évolutive.

Les questions recherche ayant été à la base de ces recherches concernaient la faisabilité d'un profilage utilisateur basé sur des ontologies et l'applicabilité d'une telle approche dans un domaine tel la publicité en ligne. Les contributions présentées ont concerné la définition d'une ontologie modulaire adaptée au

domaine de la publicité en ligne, la spécification d'un ensemble de règles logiques traduisant des connaissances métier et un processus de profilage. Le modèle d'ontologie développé prouve qu'il est possible de décrire les éléments intervenant dans un domaine d'application donné (e.g. publicité en ligne) en utilisant des techniques d'ingénierie d'ontologies. L'ontologie ainsi produite contient des concepts et des relations entre ces concepts reflétant la réalité de l'eco-système de la publicité en ligne. Ce modèle ontologique peut, en outre, être étendu avec des précisions non pas sur le domaine de la publicité en ligne, mais concernant le processus observé pour le profilage des utilisateurs du système. Les différentes règles spécifiées permettent à un raisonneur de suivre les mêmes étapes que suivrait un expert du domaine. Nous avons proposé une nouvelle approche permettant de manipuler des relations pondérées au travers de techniques d'ingénierie d'ontologies traditionnelles. La correctitude des résultats produits par le prototype MindMinings a été vérifiée à travers différentes évaluations.

3.2.6.2 Ouvertures

Les résultats numériques de ces évaluations montrent qu'une structure ontologique avec un haut niveau de complexité n'est pas aisément manipulable avec des outils gratuits. Alors que les solutions logicielles choisies (OWL API, Protégé) permettent tout de même de récupérer des résultats, les temps nécessaires à l'obtention de ces résultats ne sont absolument pas optimisés. De plus, on remarque le fait que le système retourne des réponses uniquement lorsqu'un faible nombre d'utilisateurs, d'univers et de pages Web sont considérés. Il est dès lors difficilement concevable que, sans aucune adaptation, ce modèle ontologique pourra être utilisé dans un contexte avec des milliers d'utilisateurs, d'univers et de pages Web. Toutefois, des recherches menées par la suite me permettent aujourd'hui d'apporter des éléments de réponse à ce problème. En effet, il faudrait combiner cette approche avec la méthode de sélection de règles logiques (présentée dans la section 0) afin de déterminer quel effet cela aurait sur la performance globale du système. Une autre optimisation possible serait de placer l'ensemble de la base de connaissances dans un magasin de triples haute performance (e.g. Stardog avec la version commerciale du raisonneur Pellet). Enfin, il serait aussi intéressant d'investiguer ce qu'un changement de type de raisonneur aurait comme effets sur les durées de classification. L'approche actuelle reposant sur un chaînage arrière, un raisonneur chaînage avant, e.g. RuQAR (Bak et al. 2014) ou BaseVISor (Matheus et al. 2006), couplé à un changement du langage utilisé pour la description de l'ontologie (passer du OWL DL à du OWL 2 RL par exemple) pourraient aider à diminuer les temps nécessaires à la classification des pages Web et au profilage d'utilisateurs.

Concernant la comparaison avec les approches à base d'algorithmes d'apprentissage automatique, l'approche présentée ici a des résultats bien meilleurs en termes d'exactitude et de correctitude des résultats produits. Ceci se justifie principalement par la dépendance à un domaine des techniques d'apprentissage automatique. Or, par définition, le Web est "multi-domaines". Alors qu'un système entraîné sur un univers donné (par exemple la puériculture) pouvait arriver à de bons résultats, dès qu'il était appliqué à des pages Web d'un autre domaine (par exemple Politique), son niveau de précision chutait. Ce comportement peut aussi s'expliquer par le fait que, par rapport aux différents univers identifiés, les différents corpus d'apprentissage constitués par l'entreprise n'avaient pas une distribution homogène. Dans tous les cas, notre approche s'est révélée plus performante en termes de précision et de facilité d'utilisation par rapport à l'approche à base d'apprentissage automatique.

De nos jours, le modèle utilisé pour la publicité en ligne est en train d'évoluer. Le modèle actuel a atteint ses limites - les exemples de sites tels Facebook ou YouTube le confirment. Le meilleur moyen de gagner de l'argent avec le modèle publicitaire du Web 2.0 revient souvent à seulement diffuser des annonces sur un site générant beaucoup de trafic. Par contre ces publicités ne sont pas profilées pour un type d'utilisateur en particulier. Cela résulte en un nombre de plus en plus élevé de personnes qui utilisent des bloqueurs de publicité - même Google a une telle extension dans son navigateur Chrome. Avec l'avènement des technologies blockchain, cela est en train de changer. Le créateur du langage JavaScript (et fondateur de Mozilla), Brendan Eich, a lancé le projet "*basic attention token*[17]". Basé sur la blockchain Ethereum, l'approche consiste à utiliser un jeton à échanger entre éditeurs, annonceurs et internautes. C'est le navigateur utilisé, Brave, qui étudie comment l'internaute passe son temps et génère les jetons nécessaires pour récompenser les éditeurs ayant capté l'attention de l'utilisateur. Ce service crée un marché de la publicité numérique basé sur une approche blockchain transparente et efficace. Les éditeurs reçoivent plus de revenus parce que les intermédiaires et la fraude sont réduits. Les utilisateurs, qui y

[17] https://basicattentiontoken.org

participent, reçoivent des annonces moins nombreuses mais mieux ciblées, moins exposées aux logiciels malveillants. Et les annonceurs obtiennent de meilleures données sur leurs dépenses. Il ne s'agit pas de l'unique initiative en rupture avec le modèle en place - Oster Pearls[18] permet aux consommateurs de contenus de contribuer aux sites Web qu'ils aiment, en permettant qu'une petite portion de leur CPU et GPU soit utilisée pour du stockage de fichiers anonyme et décentralisé, rendant dès lors les sites moins dépendants des publicités.

Même si les travaux présentés ici n'ont pas donné lieu à l'implémentation escomptée, ils ont marqué le début de ma double interrogation concernant l'impact de la modélisation sur les performances de l'implémentation et vice versa. En effet, dans les recherches présentées ici, nous avons exploré les possibilités offertes par les ontologies et les langages de règles, sans nous soucier des contraintes d'implémentation. Les différentes évaluations quantitatives du système ont mis en évidence le fait que même si les résultats de la classification ou du profilage étaient corrects, les temps d'exécution associés ne permettaient pas une implémentation telle quelle de l'approche. Dans la suite de mes recherches, j'ai essayé de toujours trouver un équilibre entre la modélisation et l'implémentation. D'une part, lorsque l'implémentation est moins importante, il est possible d'argumenter en faveur de modèles ontologiques construits proprement, c'est-à-dire en respectant les standards et bonnes pratiques. Ce fut notamment le cas pour la conception de l'ontologie ifcWoD (section 3.3.4.2.3). D'autre part, lorsqu'il est important de respecter des contraintes d'implémentation, la modélisation peut être adaptée afin de les intégrer. Comme présenté pour l'approche FOWLA (section 3.3.4.2.4), sans le module de sélection de règles, il aurait été impossible de proposer une fédération d'ontologies respectant les contraintes définies pour les temps d'exécution de requêtes.

[18] https://oysterprotocol.com

3.3 Fédération d'ontologies à base de règles

3.3.1 Perspective historique

Ce chapitre présente les recherches menées au cours du co-encadrement des travaux de thèse de M. Tarcisio MENDES de FARIAS, traitant de l'interopérabilité sémantique des systèmes d'information d'entreprise. C'est à la suite de la signature d'un contrat de thèse CIFRE N°2013-1200 avec la société Active3D (Dijon, France), et sous la direction du Prof. Christophe Nicolle, que j'ai co-encadré les travaux de doctorat de Monsieur MENDES de FARIAS. Cette thèse fut soutenue en octobre 2016. Monsieur MENDES de FARIAS a, par la suite, signé un contrat CDI avec l'entreprise Dassault Systems à Aix en Provence. Au bout de 6 mois, se voyant affecter que des tâches liées à l'implémentation du format IFC dans les logiciels de l'employeur, Monsieur MENDES de FARIAS a choisi de démissionner. Il effectue depuis un contrat postdoctoral à l'Université de Lausanne, dans l'équipe du Prof. Dessimoz. Nous continuons d'échanger et de collaborer activement - nous venons de publier un article de journal au début de cette année 2018 [RIS1].

Les travaux présentés ici s'inscrivent dans le cadre de la recherche en fédération d'ontologies pour améliorer l'exploitation d'expertises métiers dans les domaines de l'architecture, de l'ingénierie et de la construction. Ces travaux, comme ceux du chapitre précédent, ont reçu une forte impulsion de départ de la part de l'entreprise partenaire, et ont été contraints par les différents standards et approches standard utilisés par les experts du domaine. C'est pour cela, qu'avant de présenter les contributions scientifiques et surtout afin de bien cerner la problématique associée aux recherches menées dans ce contexte, les paragraphes suivants vont présenter l'écosystème BIM (*Building Information Modeling*), tel que défini par les différents organismes de standardisation.

La dématérialisation de l'ensemble des données et des processus dans le domaine du bâtiment est un enjeu mondial depuis ces 10 dernières années. L'idée est de pouvoir représenter de manière homogène l'ensemble des données produites tout au long du cycle de vie d'un bâtiment, et que chaque acteur puisse les utiliser et les enrichir. Or, un grand nombre d'acteurs est impliqué dans la conception, la réalisation et l'exploitation d'un ouvrage. Chaque acteur est responsable de la création d'une partie de l'information concernant l'ouvrage et utilise une partie de l'information produite par les autres comme donnée d'entrée nécessaire à l'exécution de sa tâche. Afin de répondre à ce besoin, l'approche BIM s'est imposée en tant que standard international.

Lorsque l'on cherche à définir le BIM, de nombreuses définitions apparaissent (Volk et al. 2014). Pour la suite des discussions, je vais reprendre la définition donnée par l'ISO, à savoir le BIM concerne l'utilisation d'une "représentation numérique partagée d'un objet construit (comprenant bâtiments, ponts, routes, usines, etc.) pour faciliter les processus de conception, de construction et d'exploitation et former une base fiable permettant les prises de décision" (ISO29481-1 2016).

Le BIM est d'abord un modèle entités-relations (approche objet, relations, attributs) décrivant les données échangées remplaçant le traditionnel paquet de documents (plans et pièces écrites). Outre la disponibilité de nouvelles possibilités de visualisation de la géométrie, l'objectif est aussi de déléguer une partie des contrôles de cohérence des données échangées à la machine. On ne décrit plus des documents, dont la coordination est à la charge de l'utilisateur, mais des composants du produit final. On complète aussi le modèle avec les "non-dits" que l'humain sait traiter, en rajoutant de la sémantique et des relations entre les entités.

Le BIM est aussi une méthode décrivant comment renseigner le modèle de données selon les étapes du cycle de vie, afin qu'il soit exploitable par l'homme et la machine, et qu'il ne contienne, à un instant donné du cycle de vie, que les données pertinentes, et seulement celles-ci. Cette nouvelle approche de

l'information échangée ne remet pas en cause l'organisation actuelle en corps de métier, leurs rôles respectifs, leurs responsabilités, ainsi que la protection de leur savoir-faire. L'approche ne concerne que le support et l'exploitation de l'information échangée. Elle couvre tout le cycle de vie de l'ouvrage, de sa conception à sa démolition.

La première étape de normalisation du BIM a été réalisée en 1999 par l'IAI (maintenant buildingSmart International) (Eastman et al. 2011). Cette organisation a défini plusieurs standards ouverts, afin d'encourager le développement et l'implémentation par les acteurs du domaine d'approches dites "*open BIM*", c'est-à-dire des approches utilisant ces différents standards.

Le BIM s'appuie sur l'ingénierie des systèmes, qui se focalise sur la définition des besoins du client et des exigences qui en découlent, avec trois points de vue complémentaires, chacun implémentant un standard différent:

> Opérationnel ou "pourquoi" - consiste en des exigences réglementaires ou des besoins clients, et est exprimé à travers des MVD (*Model View Definition*)
> Fonctionnel ou "quoi" - consiste en des cas d'usage BIM et spécifié dans des IDM (*Information Delivery Manual*)
> Organique ou "comment" - spécifié via des fichiers IFC (*Industry Foundation Classes*)

Afin de mieux comprendre les interactions et relations entre ces standards, je vais utiliser un schéma de la norme (ISO8000-110 2009) pour illustrer la manière dont on doit comprendre l'environnement normalisé des échanges d'informations dans un contexte BIM.

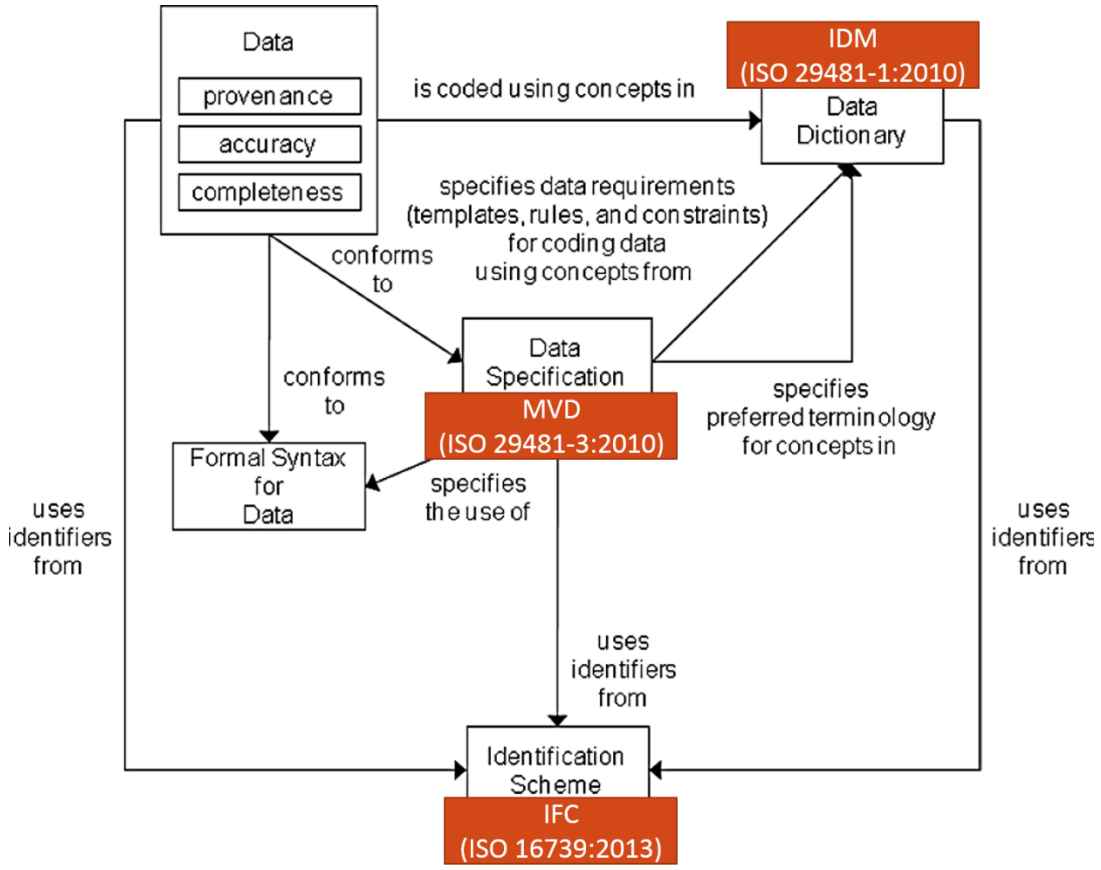


Figure 40. Environnement normalisé des échanges d'informations BIM (ISO8000-110 2009)

Ainsi, les trois standards apparaissent clairement en tant qu'outils complémentaires pour la spécification d'échanges entre acteurs:

> L'IDM spécifie les informations échangées dans le cadre de processus de référence. Il comprend une cartographie de ces processus et une ou plusieurs spécifications d'échange exprimées dans des termes compréhensibles par des utilisateurs et non ambigus. Dans les normes ISO qui traitent du sujet

(ISO29481-1 2016), "*Information Delivery Manual*" est traduit par "Contrat d'Interchange". L'IDM est une description en langage naturelle de l'échange.

- Le MVD décrit le modèle de données nécessaire pour satisfaire les besoins d'échange. En d'autres termes, un MVD est la traduction informatique d'un IDM. Comme pour les IDM, la méthodologie de description d'un MVD est normalisée (ISO29481-3 2010). Un MVD permet d'extraire un sous-ensemble du modèle IFC tel que réalisé par l'outil métier selon les indications contenus dans l'IDM. buildingSMART International (bSI) a défini le format MVDXML dont les spécifications sont publiques (BuildingSMART 2016). Ce format permet de décrire un MVD et les spécifications des échanges associées. Il répond à divers usages :
  - La définition d'un sous-ensemble des IFC
  - La validation automatique de jeux de données IFC, dans le cadre de contrôles qualité et de la certification des logiciels
  - La génération de la documentation du modèle IFC, complet ou d'un sous ensemble.
  - Aider les éditeurs de logiciels à proposer des filtrages de données IFC basés sur des MVD
- Enfin, l'IFC (Liebich et al. 2013) représente le modèle conceptuel des données standard (ISO16739 2013) et est utilisé pour la représentation de l'ensemble des éléments constitutifs du bâtiment physique. Le modèle IFC est spécifié à la base en langage EXPRESS, conforme à la norme (ISO10303-21 2002) ou *STEP part 21* (*STandard for the Exchange of Product model data*). STEP porte sur la représentation et l'échange de données de produits et a pour objectif d'intégrer les processus de conception, de développement, de fabrication et de maintenance de ces derniers. Contrairement à d'autres formats, tels le DXF (*Drawing eXchange Format*) (Autodesk 2010) ou le DWG (*DraWinG*) (Alliance 2013), orientés sur une représentation graphique et vectorielle, la norme IFC est basée sur une modélisation objet. Cette norme permet l'interopérabilité des outils et des processus autour de la maquette numérique. Grâce à cette représentation orientée objet, l'IFC permet d'identifier de manière unique chaque élément au sein d'une maquette numérique donné par un identifiant global (appelé GUID pour *Globally Unique IDentifier*) et d'associer les éléments les uns aux autres sous la forme d'un graphe. Le modèle IFC a l'ambition de couvrir tout le cycle de vie. Par conséquent il est très riche afin de pouvoir s'adapter à l'évolution et permettre l'enrichissement des informations échangées.

Toutefois, les différents processus pouvant être implémentés sur la base de ces standards dépendent encore grandement d'opérateurs humains. De plus, les différentes évolutions subies par le modèle IFC (notamment l'intégration d'ajouts pour décrire des ponts e.g. IfcBridge, des infrastructures routières e.g. IfcRoad et ferroviaires e.g. IfcRail) n'ont pas aidé à le rendre plus compréhensible ni à faciliter la manipulation des fichiers IFC par les professionels du domaine. En effet, le principal défi associé à la mise en oeuvre du BIM concerne l'équilibre à atteindre entre flexibilité du modèle IFC et rigueur de la description des informations échangées.

Cet écosystème BIM soulève en effet plusieurs problèmes, dont les principaux sont listés ci-dessous:

- **Grande raideur** - Le BIM se veut une approche collaborative pour l'ensemble des acteurs d'un projet de construction, dans laquelle chaque acteur modélise ce qui le concerne, puis contribue au modèle projet. Afin que ces différentes contributions soient correctement interprétées et donc intégrées, il faut que les différents échanges d'informations soient rigoureusement décrits. Or, il n'en est rien: très peu d'exemples utiles d'IDM existent, les MVD définis par buildingSmart sont bien trop généraux et ne permettent pas un filtrage efficace des IFC. De plus, la typologie très variées d'acteurs et de projets n'arrange rien.
- **Extraction, modification et agrégation de fichiers IFC** - Contrairement à ce que son laisse penser, les GUIDs ne sont uniques qu'au sein d'une même maquette numérique, donc au sein d'un seul fichier. Cela soulève des problèmes au niveau de l'intégration de plusieurs fichiers IFC en un seul - par exemple lorsqu'on souhaiterait combiner plusieurs point de vue (fichiers IFC correspondant à différents MVDs) en un seul fichier IFC. De plus, les attributs graphiques (couleur, texture...) n'ont pas d'attribut GUID. Il faut donc soit les redéfinir à chaque usage (ce qui peut nuire à une certaine homogénéité de la représentation graphique), soit les adresser à plusieurs objets différents sans utiliser une relation ayant un attribut GUID (ce qui pose des problèmes lors de l'agrégation de plusieurs modèles). C'est aussi le cas pour certaines entités géométriques de base (e.g. l'origine d'un repère ou des vecteurs unitaires) qui, selon le standard IFC, sont définies une seule fois puis simplement invoquées à la volée. En pratique, cela revient à invoquer des entités sans GUID, ce qui

s'accompagne de risques d'erreurs non-négligeables lors d'agrégation ou de modification de modèles.

> **Structure** - Le modèle IFC permet de décrire la hiérarchie des espaces dans l'ordre projet - site - bâtiment - étage - pièce. Mais elle ne décrit pas directement les relations entre les repères locaux de ces espaces, pourtant fondamental pour le positionnement des composants. Il n'y a pas non plus de relation directe entre une annotation et l'élément qu'elle enrichit. Or, au sens IFC, une annotation est une représentation graphique dans un contexte géométrique et qui apporte une information aux objets constituant le modèle. Une annotation est un élément graphique additionnel comme par exemple un point, une courbe, un texte, une cote, des hachures, etc. La seule relation spécifiée par le modèle IFC est une relation de positionnement dans un espace (`IfcRelContainedInSpatialStructure`). Ainsi, il n'est pas possible d'exploiter ces annotations autrement que visuellement.
> **Trop permissif** - L'utilisation du format IFC est freinée par la richesse du modèle qui n'est pas canalisée et permet la permissivité. La norme IFC n'étant pas assez stricte, un même composant peut être modélisé de plusieurs manières différentes. En outre, des divergences existent au niveau des implémentations logicielles (les développeurs de logiciels ne font pas tous les mêmes choix). L'import d'un fichier IFC dans un logiciel a donc de grandes chances de générer des erreurs si ce fichier IFC a été conçu ou exporté à partir d'un logiciel différent.
> **Pas de lien vers les SIG** - Aujourd'hui la description du bâtiment se réduit à ce qui est construit. Les fondations flottent "en l'air". Les fouilles ou les remblais sont ignorés. Il serait important de disposer d'un élément volumique de sol permettant de traiter la problématique du métier de terrassier ou de spécialiste de fondations spéciales. Les réseaux connectés à l'ouvrage devraient suivre le même principe.
> **Pas de prise en compte de l'évolution temporelle** - Un objet se caractérise par des représentations géométriques et des propriétés géométriques et/ou non géométriques. Les représentations géométriques sont plurielles car elles peuvent dépendre de points de vue différents. Il n'y a pas nécessairement de lien entre les représentations géométriques et les propriétés géométriques. Cette permissivité apparente permet de gérer par exemple les niveaux de "détail/développement". Mais aujourd'hui les choix possibles ne sont pas encadrés par la norme. Et cela est un frein à la mise en oeuvre du BIM, car le résultat est dépendant du choix de l'éditeur de logiciel.
> **Sémantique** - Pour préserver la qualité et l'unicité du vocabulaire décrivant les entités IFC, bSI a entrepris de mettre tous les mots dans bSDD[19] (*buildingSMART Data Dictionnary*). Et cela concerne aussi les mots définissant les propriétés. Le WG4 du CEN/TC 442 animé par la France est moteur sur ce sujet avec la norme AFNOR NF XP P07-150. Il n'y a aucun lien défini entre le vocabulaire IFC et les différentes initiatives de structurations des termes en vocabulaires (sémantiques ou pas). Or ce travail est fondamental si l'on souhaite éviter la trop grande permissivité actuellement existante dans les échanges d'informations basées sur le format IFC.

D'autres éléments pourraient être ajoutés à cette liste, comme les erreurs d'importation des fichiers IFC (lorsque les attributs non supportés sont tout simplement perdus). Ce qui est sûr c'est que le BIM est devenu obligatoire pour les marchés publics en France en 2017, alors qu'il l'était déjà dans d'autres pays (depuis 2015 en Grande-Bretagne, depuis 2007 en Finlande). L'idée n'est donc pas d'énumérer les problèmes avec les standards actuels, mais bien d'identifier des approches permettant de faciliter la vie des professionnels du domaine et des éditeurs logiciels. Sur cette base, la section suivante discute d'approches issues du domaine de l'interopérabilité des systèmes d'information, à partir desquelles nous avons construit, en partie, nos contributions dans ce domaine.

3.3.2 Description du domaine et des problèmes étudiés

Allant au-delà de la simple représentation 3D d'un bâtiment, le BIM peut être vu comme un système d'information coopératif complexe, intégrant avec difficulté les données et pratiques hétérogènes des différents acteurs du domaine. Pour répondre efficacement aux différents défis sous-jacents, discutés dans la section précédente, l'idée est ici de s'inspirer des approches existantes dans le domaine de l'EII (*Enterprise Integration and Interoperability*). Ce domaine traite justement des problèmes liés à l'absence d'interopérabilité entre Systèmes d'Information (SI) dans les organisations, et c'est le principal problème sous-jacent à l'implémentation du BIM.

[19] http://bsdd.buildingsmart.org

Lorsque l'on considère les besoins en gestion de données, de connaissances et face à la nécessité de devoir réagir rapidement aux changements des forces du marché, il est évident que la collaboration entre acteurs BIM ne peut reposer que sur des technologies de l'information. Améliorer (voire même maximiser) cette collaboration implique l'application des principes sous-jacents à l'intégration et à l'interopérabilité des entreprises.

L'EII a fait son apparition vers la fin des années '90, juste après l'avènement du World Wide Web (WWW). En règle générale, l'interopérabilité sous-entend les notions d'autonomie, d'environnement fédéré ou encore coexistence; l'intégration désigne les notions de coordination, d'uniformité ou encore cohérence (Chen et al. 2008), (H. Panetto & Cecil 2013). Dans un contexte industriel, les approches EII visent à fournir des outils permettant l'intégration de données hétérogènes, provenant de diverses sources, sans avoir besoin de toutes les charger, au préalable, dans un entrepôt central (Halevy et al. 2005). Dans un contexte recherche, nous pouvons séparer cette discipline en deux domaines de recherche: (i) les technologies de l'information et (ii) la modélisation d'entreprise (*enterprise modeling*). Les technologies de l'information sont étroitement liées au contexte de l'industrie, où les outils développés font référence à des systèmes d'intégration de données. La modélisation d'entreprise fait référence à un ensemble d'approches et de concepts, comprenant la conception d'architectures système globales, la gestion de la cohérence entre objectifs locaux et globaux, l'allocation des ressources ou encore la cohérence des données (Vernadat 2002).

Typiquement un modèle d'entreprise se compose de données structurées et de règles relatives aux informations que les applications utilisent pour réaliser leurs tâches. Étant donné la taille et la complexité des entreprises, il est difficilement imaginable qu'un modèle unique pourra décrire ce qui s'y passe à tout moment. En raison de la diversité des applications dans une entreprise, les informations contenues dans les modèles sont codées à travers de nombreux formats selon les besoins des émetteurs et des consommateurs d'informations. Les modèles d'entreprise peuvent prendre diverses formes, allant d'organigrammes à des données de flux de processus, en passant par des feuilles de calcul, des dessins techniques, ou encore des fichiers de fabrication assistée par ordinateur. Tous ces exemples représentent des îlots d'information ou d'automatisation, à la manière de représentations de connaissances non connectées entre elles.

En effet, d'un point de vue utilisateur ce qui importe c'est que ces différents modèles soient portables, à savoir pouvoir être réutilisés d'une application à l'autre, sans dépendre de la configuration spécifique de l'une d'entre elles. La solution qui se profile semble être un modèle unifié ou intégré, qui fournirait un modèle d'information unique quelle que soit l'application. Une telle approche améliorerait la capacité des utilisateurs à fournir des informations provenant de modèles hétérogènes pour être utilisées par des applications et des plates-formes au niveau de l'entreprise. La définition de modèles standard pourrait aider à atteindre cette vision. Toutefois, le problème d'hétérogénéité des représentations demeure: il y aura toujours autant de manières différentes de présenter un modèle qu'il y aura des raisons de vouloir modéliser le domaine d'intérêt.

En conséquence, à partir de la fin des années '90, la tendance au niveau des EII fut de s'intéresser de plus en plus à l'interopérabilité des entreprises (ou des modèles d'entreprises). De nombreuses définitions existent pour l'interopérabilité (Vernadat 1996), (Chen & Vernadat 2003), (Chen & Vernadat 2004). Par exemple, dans (Vernadat 1996), l'auteur définit l'interopérabilité comme la capacité de communiquer avec des systèmes pairs et d'accéder aux fonctionnalités de ces systèmes. L'ISO définit l'interopérabilité comme la "capacité que possèdent des systèmes à fournir des services et à recevoir des services d'autres systèmes et à utiliser les services ainsi échangés afin de fonctionner ensemble de manière efficace" (ISO17261 2012). D'un point de vue machine, implémenter une interopérabilité revient à relier deux systèmes informatiques hétérogènes, afin qu'ils puissent collaborer, ce qui implique un accès réciproque à leurs ressources.

C'est ainsi que l'interopérabilité d'entreprises (*enterprise interoperability*) ou IE, et plus particulièrement l'interopérabilité des systèmes d'information des entreprises, ont été reconnues ces dernières années comme d'une grande importance à la fois dans l'industrie, mais aussi dans la recherche (Charalabidis et al. 2010). Définie comme la capacité d'échanger des informations et des services entre des systèmes d'entreprise (Chen et al. 2008), l'IE vise à concevoir des approches répondant aux besoins en interopérabilité des entreprises, approches spécifiquement conçues pour une utilisation au sein de

systèmes décentralisés, dynamiques et interconnectés. A cet égard, les approches IE sont plus rentables et plus rapides à mettre en oeuvre que des approches traditionnelles.

D'un point de vue strictement informatique, l'interopérabilité peut être implémentée à trois niveaux:

- ***Physique*** - ce niveau d'interopérabilité a été résolu à travers des standards *hardware* (e.g. Ethernet) et les protocoles des couches basses de l'architecture réseaux ISO/OSI (e.g. TCP/IP et HTTP)
- ***Syntaxique*** - ce niveau concerne la syntaxe des messages échangés entre ordinateurs et a été résolu à travers le langage XML et les standards de syntaxe (e.g. HTML, WSDL, SOAP)
- ***Sémantique*** - ce niveau adresse le sens des messages ou ressources manipulées (e.g. pages Web) par les ordinateurs, et n'a pas encore été résolu. Les standards du Web sémantique sont un moyen de spécifier le sens d'une ressource informatique afin que des traitements automatiques des informations puissent être implémentés. Il existe différents degrés d'interopérabilité sémantique, selon les différents langages existants pour la description d'ontologies:
  - ***Minime***: Ce degré est activé par l'utilisation de RDF et correspond à la connaissance minimale pouvant être véhiculée au travers d'un énoncé (à savoir ce qui est exprimé dans l'énoncé lui-même). L'objet "*Paris*" est relié à l'objet "*France*" par la propriété "*est la capitale de*".
  - ***Etendu***: L'interopérabilité sémantique étendue correspond à l'ensemble minimal de croyances sur la base duquel, deux agents informatiques, après avoir échangé un énoncé, vont pouvoir faire des déductions. Ce niveau est permis par le langage RDF Schema. Les entités dans un énoncé étant identifiées par leurs URIs, des agents informatiques peuvent accéder à une ontologie partagée précisant en RDFS, par exemple, qu'une capitale est une ville, que les capitales sont uniques, etc. Il y a entente sur ce que sont les entités de l'énoncé, il n'y a pas entente sur ce qu'ils ne peuvent pas être (on spécifie une connaissance avec une borne minimum, mais pas de borne maximum)
  - ***Total***: L'interopérabilité sémantique totale est activée au travers du langage OWL, puisqu'il permet de spécifier à la fois ce qui existe que ce qui ne peut pas exister. Par rapport au niveau précédent, cela revient à avoir une entente entre agents informatiques avec une borne basse (ce que les agents peuvent croire) et une borne haute (ce que les agents ne peuvent pas croire). Si l'on reprend l'exemple précédent, cela revient à avoir une ontologie OWL partagée par les agents informatiques échangeant des énoncés. Cette ontologie pourra leur spécifier qu'on ne peut pas croire que "*Dijon*" est aussi capitale de la "*France*".

Partant de ces considérations, et en repensant à la vision idéale du BIM, impliquant collaboration sans coutures entre acteurs, le verrou identifié concerne ***l'interopérabilité sémantique***. Et c'est justement ce niveau d'interopérabilité qui serait souhaité pour l'écosystème BIM. Notre problématique a été donc de déterminer si les technologies du Web sémantique peuvent supporter ce type d'interopérabilité et si oui, avec quel niveau de performance. Les contributions présentées dans ce chapitre ont été motivées par les questions de recherche suivantes:

- Évaluer la pertinence des approches automatiques pour générer l'ontologie ifcOWL
- Est-il possible de modéliser d'autres standards BIM avec des langages de description d'ontologies standards ?
- Déterminer s'il est possible d'implémenter une interopérabilité sémantique entre modèles BIM sur la base d'ontologies ?
- Est-il possible d'inférer de nouveaux alignements entre les ontologies considérées ?
- Est-il possible d'intégrer des politiques d'accès aux données lors de l'exécution de requêtes, mais aussi au niveau des ontologies elles-mêmes ?
- Évaluer l'apport des alignements d'ontologies dans l'interopérabilité sémantique
- Évaluer la performance d'une telle approche pour l'exécution de requêtes
- Déterminer l'applicabilité d'une telle approche à l'écosystème BIM
- Est-ce qu'une approche à base d'ontologies permet de répondre à toutes les contraintes du domaine?
- Quel est le niveau de performance atteint par notre approche en comparaison avec les autres approches existantes ?

Cela peut d'ailleurs être expliqué le grand engouement dans la communauté BIM pour les technologies du Web sémantique et plus particulièrement l'approche des données liées. C'est aussi à ce niveau que se situent nos contributions. Toutefois, avant de les présenter, il est important de comprendre comment elles

se situent par rapport à l'existant. C'est pour cette raison que la section suivante contient un résumé des différentes approches employant les technologies sémantiques et appliquées au BIM.

3.3.3 Rappel de l'état de l'art spécifique à la thématique considérée

Pour le gestionnaire de patrimoine, le BIM prend forme à partir d'une maquette numérique à laquelle s'additionnent toutes les données attributaires définies lors de la phase de conception-construction, mais aussi en relation avec les processus métiers propres au fonctionnement de l'entreprise ou de l'administration propriétaire du bien immobilier. Ceci laisse présager une hétérogénéité au niveau des données, des processus, des acteurs ainsi que des systèmes.

En effet, au cours des deux dernières décennies, l'industrie de l'architecture, de l'ingénierie et de la construction (*Architecture, Engineering, Construction* ou AEC en anglais) a vu apparaître une multitude de SI BIM. Parmi ceux-ci, on peut citer: Revit BIM d'Autodesk, BimSight de Tekla ou encore la solution Active3D (développée par l'entreprise partenaire de ce contrat de recherche). Comme c'est le cas en général en informatique, plus il y a de logiciels, plus la collaboration entre utilisateurs est difficile. Et le format d'échange standard, IFC, n'a pas beaucoup aidé (voir discussion dans la section précédente). Apporter de l'interopérabilité sémantique dans ce contexte passe par le besoin de représenter des données bâtiment avec des technologies sémantiques (Eastman et al. 2008) et de permettre aux différents systèmes d'information BIM de manipuler ce type de données.

Ainsi, plusieurs initiatives visant à traduire le modèle IFC en langage OWL ont vu le jour dans les quinze dernières années (Schevers & Drogemuller 2005), (Beetz et al. 2009), (Zhang & Issa 2011), (Gao et al. 2015). L'ensemble des approches présentées dans la littérature repose, d'une part, sur une traduction en langage OWL de la norme IFC, et, d'autre part, sur la conception d'ontologies modélisant les connaissances métiers du domaine d'application considéré (gestion de patrimoine, conception architecturale, etc.)

- Développée dans le cadre du projet IntelliGrid (Gehre et al. 2006), l'approche décrite par (Beetz et al. 2009) est l'une des plus utilisées pour traduire la norme IFC en langage OWL. Les auteurs y présentent deux méthodes automatiques de construction d'une ontologie OWL de la norme IFC : l'une utilise le schéma XML, l'autre exploite le schéma EXPRESS de la norme IFC. La dernière méthode offre l'avantage de traduire, sans perte d'expressivité, l'ensemble des règles WHERE présentes dans le modèle IFC. Toutefois, cette traduction a été définie pour la version IFC 2x2 et ne couvre qu'une partie (850 classes) des constructions IFC.
- (Dibley 2011) présente une méthode de construction d'une ontologie IFC basée sur le parcours de fichiers STEP contenant des schémas IFC, à l'aide de la librairie OpenIfcJavaToolbox[20]. Malheureusement, la traduction en langage OWL ne porte que sur un ensemble restreint de classes IFC, celles considérées comme pertinentes dans le cadre du projet considéré (implémentation d'un environnement informé pour agents).
- (Zhang & Issa 2011) est une autre approche pour construire une ontologie OWL de la norme IFC 2x3. L'idée sous-jacente est de faciliter la récupération d'informations à partir d'un modèle IFC. Les auteurs détaillent les étapes du développement de l'ontologie, sans pour autant proposer des règles logiques permettant l'adressage de requêtes au format XML. Les auteurs discutent les avantages qu'une extension de l'ontologie IFC pourrait apporter (notamment la possibilité d'atteindre plus d'expressivité dans la formulation des requêtes), mais ne spécifient pas les concepts ou la structure de cette ontologie étendue. Cette approche choisit une méthode manuelle pour la construction de l'ontologie OWL à partir des fichiers IFC.

Les travaux présentés portent souvent sur une partie seulement du modèle IFC et proposent des méthodes de construction automatique. A partir des différentes procédures de conversion publiées, les discussions du groupe de travail Linked Data de bSI (auxquelles j'ai participé) ont permis de concevoir une ontologie communément reconnue au sein de bSI comme la traduction en OWL de l'IFC. Appelée ifcOWL[21], elle a été présentée pour la première fois au sommet bSI de Singapour, en septembre 2015.

[20] http://www.openifctools.org
[21] http://openbimstandards.org/standards/ifcowl/

Cette ontologie respecte les trois règles suivantes (par ordre d'importance) :

> L'ontologie ifcOWL doit être exprimée en OWL 2 DL
> L'ontologie ifcOWL doit correspondre en tout point au schéma EXPRESS. Ce choix a été fait, principalement, pour favoriser la reconnaissance de cette approche par la communauté bSI
> L'ontologie ifcOWL a pour principal but de convertir des données au format IFC/STEP en des triples RDF équivalents. Deuxièmement, elle donne la possibilité d'instancier l'ontologie ifcOWL sans devoir passer par la conversion des données STEP en RDF.

La procédure de conversion, d'EXPRESS vers OWL, utilisée par les auteurs est présentée dans (Pauwels & Terkaj 2016). Elle est résumée dans le Tableau 20. Il est à noter que, parmi les différents algorithmes de conversion entre EXPRESS et RDF e.g. (Krima et al. 2009), (Barbau et al. 2012), (Hoang & Törma 2014), celui utilisé par les auteurs de (Pauwels & Terkaj 2016) est le seul qui permet de générer une ontologie en OWL 2 DL (niveau d'expressivité $\mathcal{SROIQ}(\mathcal{D})$) et aussi le seul à générer des propriétés inverses. Une comparaison entre les métriques des différentes ontologies, générées avec chacun des trois convertisseurs existants, permet de démontrer ces affirmations (voir Tableau 21).

Concepts en EXPRESS	**Concepts en OWL**
SCHEMA	ontologie
Type de données simple	Classe OWL avec une restriction sur une **`owl:DatatypeProperty`**
Type de données défini	Classes OWL
Type de données SELECT	Classe OWL équivalente à une union de classes OWL
Type de données ENUMERATION	Classe OWL équivalente à une collection d'instances **`owl:NamedIndividual`**
Entités	Classes OWL
Attributs d'entités	Propriétés objet fonctionnelles, avec des domaines et portées spécifiés ; utilisation des restrictions **`owl:AllValuesFrom`** ; utilisation des cardinalités **`owl:qualifiedCardinality`** ou **`owl:maxQualifiedCardinality`**
Attributs d'entités spécifiés en tant que SET	Propriétés objet fonctionnelles, avec des domaines et portées spécifiés ; utilisation des restrictions **`owl:AllValuesFrom`** ; utilisation des cardinalités **`owl:minQualifiedCardinality`**, **`owl:maxQualifiedCardinality`** ou **`owl:qualifiedCardinality`**
Attributs INVERSE	Propriétés objet avec des propriétés inverses spécifiées via **`owl:inverseOf`**
Attributs DERIVE	Pas convertis
Règles WHERE	Pas converties
FUNCTION	Pas converti
RULE	Pas converti

Tableau 20. Règles de conversion d'EXPRESS vers OWL définies dans (Pauwels & Terkaj 2016)

Toutefois, les ontologies seules ne permettent pas d'amener l'interopérabilité sémantique entre systèmes ou encore entre entreprises. Alors qu'une ontologie permet de représenter des connaissances d'un domaine, il est clair qu'il y aura autant de représentations de cette connaissance que de concepteurs d'ontologies (ou modélisateurs) (Benslimane et al. 2006). Au mieux, il est possible d'implémenter une interopérabilité entre agents informatiques au sein d'une même application, éventuellement au sein d'une même organisation, mais pas à une échelle globale, c'est-à-dire entre l'ensemble des entreprises du domaine par exemple. Cela ne fait pas beaucoup de sens de concevoir une ontologie juste pour le plaisir de le faire, sans chercher s'il existe d'autres ontologies spécifiant les mêmes concepts, sans définir des liens sémantiques vers les concepts de ces autres ontologies ou encore sans publier l'ontologie afin que d'autres puissent s'y référer et l'utiliser pour décrire leurs connaissances. Si l'on revient aux principes des données liées (exposées dans la section 3.1.3), le besoin de coupler les ontologies existantes apparaît clairement.

Les ontologies étant des modèles, elles peuvent être couplées de la même manière que deux modèles. Dans ce contexte, le standard (ISO14258 1999) définit trois possibilités: deux modèles peuvent être intégrés, unifiés ou encore fédérés.

Eléments de l'ontologie	**ifcOWL avec OntoSTEP** (Krima et al. 2009), (Barbau et al. 2012)	**ifcOWL par** (Hoang & Törma 2014)	**ifcOWL par** (Pauwels & Terkaj 2016)
Axiomes	17498	11009	21306
Axiomes logiques	7971	8591	13649
Classes	1348	1556	1230
Propriétés objet	1778	854	1578
Propriétés de type de données	4	9	5
Instances	1155	1158	1627
Expressivité DL	$\mathcal{ALUHN}^{(D)}$	$\mathcal{ALCON}^{(D)}$	$\mathcal{SROIQ}^{(D)}$
Axiomes de sous-classes	4257	4991	4622
Axiomes de classes équivalentes	0	268	266
Axiomes de classes disjointes	0	2429	2429
Sous-propriétés objet	186	0	1
Propriétés inverses	0	0	94
Propriétés objet fonctionnelles	62	853	1441
Propriétés transitives	62	0	1
Domaines de propriétés objet	1592	8	1577
Portées de propriétés objet	174	6	1576
Propriétés type de données fonctionnelles	0	9	5
Domaines de propriétés de type de données	7	10	5
Portées de propriétés de type de données	4	10	5
Axiomes d'assertions de classes	1627	3	1627
Axiomes d'annotations d'assertions	5240	0	3210

Tableau 21. Comparaison des différentes versions ifcOWL générées à partir du schéma EXPRESS IFC2x4 (Pauwels & Terkaj 2016)

L'intégration de modèles nécessite une forme de modèle standard par rapport auquel les autres modèles sont interprétés. Il est important pour ces modèles standard (ou de référence) d'être aussi riches que leurs modèles constitutifs. Tous les modèles peuvent être stockés sous une forme standard avec des informations filtrées ou traduites par les applications, comme dans le cas d'IRDS (*Information resource dictionary system*) (ISO10027 1990). En variante, les modèles standards peuvent être conçus par les propriétaires de modèles constituants, comme c'est le cas dans STEP (ISO10303-21 2002). La standardisation d'un grand nombre de modèles pose d'énormes difficultés. Par conséquent, la capacité de traiter quelque chose de "plus léger" que des modèles intégrés apparaît comme nécessaire. C'est ainsi qu'à partir du modèle standard (la vue globale), des modèles constitutifs (les vues locales) sont décrits.

L'unification de modèles suppose un méta-modèle fournissant une structure à un méta-niveau, structure commune à l'ensemble des modèles constitutifs. Un tel méta-modèle permet d'établir des équivalences sémantiques entre les concepts des modèles constitutifs. Ainsi, par le biais de ce méta-modèle, chacun des modèles constitutifs peut être traduit en un autre modèle. Bien sûr, il se peut qu'il y ait une perte de sémantique lors de telles traductions. Ce sont les propriétaires des modèles constitutifs qui spécifient les sémantiques normalisées. SUMM (*Semantic Unified MetaModel*), décrit dans les documents de travail du groupe ISO/IEC JTC 1/SC 21/WG 3 (ISO/IEC 1991) est un exemple de méta-modèle pouvant servir pour l'unification de modèles.

La fédération de modèles est un scénario qui sous-entend qu'il n'existe pas d'agent pouvant imposer des exigences d'équivalence sémantique entre les différents modèles considérés. Ce type d'interopérabilité exige que les modèles soient reliés de manière dynamique, par le biais de liens sémantiques entre leurs concepts et relations. Cela est similaire à la définition d'un système utilisant une terminologie prédéterminée.

Nous nous sommes donc intéréssés aux différentes approches reposant sur des couplages d'ontologies pour implémenter l'interopérabilité sémantique entre deux SI.

L'une des approches les plus connues est l'approche MDI (*Model-Driven Interoperability*), initiée lors du projet INTEROP NoE, poussée par l'Object Management Group (OMG) et basée sur une architecture MDA (*Model-Driven Architecture*) (Kleppe et al. 2003). Le projet INTEROP NoE (EU FP6) identifiait 4 niveaux d'interopérabilité (données, service, processus et métier) pour lesquelles 3 barrières étaient définies: conceptuelle, technologique et organisationnelle. L'ensemble des architectures spécifiées dans le contexte de ce projet sont basées sur des ontologies, structurées en différents niveaux (allant de l'utilisateur à l'exécution) et correspondant aux trois barrières ci-dessus: CIM (C*omputational Independent Model*), PIM (*Platform Independent Model*) et enfin PSM (*Platform Specific Model*). Les modèles CIM correspondent à des processus métier et sont spécifiés sans aucune référence à des technologies. Ils sont par la suite traduits en des modèles PIM représentant leurs déclinaisons selon un modèle de plateforme considéré (e.g. CORBA, .NET). Enfin, ces modèles sont transformés en équivalents PSM afin d'implémenter un vrai système pouvant être exécuté par les ordinateurs. Les différentes transformations entre modèles représentent des processus statiques. D'après (Agostinho et al. 2015), cette approche est toujours en cours de développement. A part l'architecture MDSEA (*Model Driven Service Engineering Architecture*) (Ducq et al. 2012), (Bazoun et al. 2013), très peu de solutions implémentent cette approche aujourd'hui. Impliquant une ontologie de référence (ontologie de l'interopérabilité, commune à l'ensemble des modèles CIM/PIM/PSM), et proposant des outils pour vérifier la cohérence sémantique des différents modèles, cette approche illustre un couplage de modèles à la frontière entre l'intégration et l'unification.

Lorsque l'on étudie les approches d'unification d'ontologies, le manque de flexibilité est évident car elles se basent sur un méta-modèle pour définir des liens entre les modèles constitutifs. Dans un contexte EIS, c'est avant tout la maintenance d'un tel méta-modèle qui est difficile et affecte l'autonomie de l'EIS. En guise d'exemple, je peux citer l'approche GeoNis présentée dans (Stoimenov 2009) et traitant de l'intégration de sources de données SIG (*Système d'Informations Géographiques*). L'interopérabilité entre les différentes sources de données ne nécessite pas de contrôle central, mais repose sur un intergiciel (*middleware*) pour donner l'accès à ces sources. Cette plateforme a depuis été utilisée pour interopérer plusieurs EIS, comme présenté dans (Stoimenov et al. 2015). Une autre approche (Agostinho et al. 2011) étudie comment il est possible d'implémenter une interopérabilité entre EIS à long terme. Pour ce faire, les auteurs proposent de décrire les données, les sémantiques ainsi que les correspondances entre schémas sont décrites avec des 5-tuples et stockés dans une base de connaissance avec des fonctions de raisonnement. Les différents composants d'un 5-tuple (un identifiant unique, les éléments alignés, le type d'alignement, le degré d'alignement e.g. schéma ou conceptuel, et la fonction de correspondance) permet de définir un alignement entre modèles. Pour représenter et stocker des 5-tuples, cette approche utilise OWL pour définir une extension de l'ontologie *Model Traceability* (Sarraipa et al. 2007) sous la forme d'un C*ommunication Mediator* (CM). L'ontologie CM devient dès lors un méta-modèle décrivant les liens sémantiques d'équivalence entre les différentes ontologies considérées. Un tel couplage unifié de modèles ne correspond pas aux besoins en flexibilité du BIM. Il nous reste à examiner la fédération de modèles.

Considérée comme l'approche la plus intéressante pour atteindre une interopérabilité maximale entre modèles par (Hervé Panetto & Cecil 2013), la fédération de modèles peut être implémentée de deux manières différentes:

- ***Fortement couplée***: Dans ce cas, on utilise des liens forts, codés en dur au niveau des ontologies, afin de les interopérer. L'ensemble de ces liens forme une ontologie globale intégrée. Une telle ontologie globale soulève les mêmes défis pour un EIS en termes d'autonomie et de maintenance. (Chen et al. 2009) et (Corry et al. 2015) sont des exemples de telles approches.
- ***Faiblement couplée***: Il n'y a alors pas besoin de définir une ontologie globale. Le degré d'adaptabilité de ces approches est plus élevé, dans un contexte EIS. Malheureusement, l'ensemble des approches étudiées ne prenaient pas en charge l'hétérogénéité des schémas, car seules des correspondances entre concepts sont considérées. De plus, nous n'avons pas trouvé d'approches permettant l'inférence automatique de nouveaux alignements entre ontologies. Dans (Rifaieh et al. 2005), les auteurs proposent l'ontologie multi-représentations (MurO), combinant la logique de description (Baader et al. 2003) avec la logique modale (Gabbay et al. 2003). MurO se veut un ensemble d'ontologies, dépendant d'un contexte, exprimé à travers une ontologie du contexte. C'est cette ontologie du contexte qui comprend les liens sémantiques entre les concepts de différentes ontologies contenues. MurO a été spécifiquement conçue pour adresser le problème des représentations multiples au sein d'un EIS. MurO ne permet pas d'inférer de nouveaux concepts. Une autre approche (Castano et al. 2006), conçue toujours dans le cadre du projet INTEROP NoE, utilise le langage OWL pour décrire

des ressources d'information de l'entreprise sous la forme d'ontologies. De par l'hétérogénéité sémantique inhérente à chaque représentation, il était possible que plusieurs ontologies décrivent (de manière différente) les mêmes ressources. Pour pallier à ce problème, les auteurs ont proposé trois services sémantiques d'interopérabilité: le service découverte (retourne les ressources pertinentes à partir d'une requête formulée via un modèle), le service correspondance (utilise le modèle de requête pour déterminer quels modèles seront utiliser pour calculer la distance sémantique entre les ressources) et le service d'acquisition de la ressource proprement dite. Cette approche présente plusieurs inconvénients, notamment elle permet de requêter uniquement des concepts (pas de possibilité d'utiliser un motif de graphe ou encore les options SPARQL telles que ORDER, LIMIT, etc.), elle ne supporte pas l'inférence de nouveaux alignements de concepts (en suivant des relations de transitivité simples telles A->B, B ->C alors A->C) et enfin son service de correspondance doit être exécuté au moment de la requête (les alignements précédents ne sont pas stockés, et ils doivent être recalculés à chaque nouvelle requête).

De cette étude, il en résulte qu'aucune des approches existantes ne pourrait convenir à un contexte BIM. Aux raisons énoncées ci-dessus, j'aimerais ajouter le fait qu'aucune des approches étudiées n'envisage la définition et la gestion de droits d'accès aux différentes données ou ressources considérées. Or dans un contexte BIM, avoir des politiques claires d'accès aux données est essentiel afin de garantir que chaque acteur demeure propriétaire de ses données et que des données confidentielles ne sont pas échangées par inadvertance. De plus, dans un contexte EIS, la performance des implémentations des différentes approches n'est presque pas adressée (les contraintes de temps sont souvent spécifiées en des termes très approximatifs e.g. "le système doit donner une réponse en un temps raisonnable"). Dans un contexte BIM, il est clair que les utilisateurs souhaitent des solutions au moins aussi performantes que ce qui existe aujourd'hui sur le marché.

3.3.4 Approche et résultats

3.3.4.1 Approche

A partir de l'étude précédente, et d'échanges avec différents professionnels du domaine, les contributions visées dans le cadre de la thèse de Monsieur MENDES de FARIAS concernent:

- L'interopérabilité sémantique entre ontologies :
  - Conception d'une architecture fédérée faiblement couplée pour ontologies OWL – [CI12] (C1)
  - Définition d'une approche à base de règles logiques SWRL – [RIS5] (C2)
  - Evaluation des performances associées en utilisant des modèles BIM – [CI8] (C3)
- Modélisation sémantique de standards BIM :
  - Nouvelle approche pour la conversion de fichiers IFC2x3 en ontologie OWL [CI18] (C4)
  - Adaptation du modèle COBie en ontologie – COBieOWL [CI10] (C5)
  - Nouvelle ontologie ifcWoD pour pallier aux inconvénients de l'ontologie ifcOWL bSI – [CI11] (C6)
- Évaluation de l'apport des technologies sémantiques pour l'extraction de vues à partir d'un fichier IFC :
  - Nouvelle approche permettant d'extraire des sous-parties d'un fichier IFC, en exploitant l'ontologie ifcOWL – [RIS1] (C7)

Les contributions ci-dessus ne sont pas listées dans un ordre chronologique, mais par ordre d'importance. Les paragraphes ci-dessous précisent comment les travaux sous-jacents se sont articulés tout au long du co-encadrement des travaux de doctorat de Monsieur MENDES de FARIAS.

Premièrement, après avoir étudié les différentes versions existantes de convertisseurs du standard IFC en ontologies, nous avons décidé d'en créer une nouvelle en basant notre travail sur les leçons tirées de nos analyses (C4). Nous avons ensuite comparé l'ontologie ainsi générée avec celle officiellement reconnue au niveau de bSI, et montré les avantages de notre approche.

Devant des résultats encourageants, nous avons poussé notre réflexion vers un autre standard du BIM, notamment l'approche COBie. Permettant de manipuler des informations du cycle de vie d'un bâtiment sous la forme de tableaux de type Excel, les fichiers COBie sont très utilisés par les gestionnaires de patrimoine. Sous l'impulsion de l'entreprise partenaire, nous nous sommes intéréssés à la "traduction" de

fichiers IFC en fichiers COBie. Cela nous a menés vers la conception (semi-automatique) d'une ontologie COBie (C5) et à la définition de règles de "traduction" entre concepts IFC et concepts COBie. Ceci nous a fait réfléchir davantage à l'ontologie ifcOWL obtenue avec notre convertisseur. Nous avons modifié manuellement la structure de cette ontologie (notamment en ajoutant de nouvelles propriétés) et fait le lien avec d'autres vocabulaires sémantiques. Nous avons ainsi conçu l'ontologie ifcWoD (Web of Data) dont nous avons montré les différents avantages par rapport à l'approche bSI (C6).

Ces travaux nous ont permis de converger vers la définition d'une architecture pour la fédération faiblement couplée d'ontologies OWL (C1). Nous avons spécifié cette architecture avec un haut niveau d'abstraction, puis nous avons évalué les performances ainsi atteintes avec des ontologies issues du BIM (C3). Nous avons ainsi déterminé le fait que les alignements à base de règles SWRL ralentissaient considérablement l'exécution de requêtes SPARQL au-dessus des ontologies alignées. C'est là que m'est venue l'idée de limiter les règles SWRL à considérer au moment de l'exécution d'une requête aux seules règles pertinentes par rapport à la requête (C2). Cela nous a permis de concevoir un module sélecteur de règles, que nous avons intégré dans notre architecture de fédération d'ontologies. Nous avons démontré que cela permettait d'améliorer grandement le temps d'exécution de requêtes, tout en retournant le bon nombre de résultats.

Enfin, pour répondre à un problème souvent rencontré lors de la manipulation de fichiers IFC avec les différents logiciels existants, et qui présentait un intérêt particulier pour l'entreprise partenaire, nous avons investigué les éventuels apports des technologies sémantiques pour l'extraction de vues IFC. La restitution intégrée de données aux usagers du bâtiment est une opération délicate. En effet, pouvoir extraire des sous-parties d'un fichier IFC présente des avantages pour les gestionnaires de patrimoine notamment. Ils pourraient ainsi construire des "vues" correspondant à un besoin spécifique, par exemple pour transférer l'enveloppe du bâtiment à des candidats d'un appel d'offres. Le fait de pouvoir extraire cette "vue" en tant que sous-partie du fichier IFC complet permet d'éviter l'échange entre acteurs de descriptions complètes des bâtiments, qui peut être néfaste pour des raisons de confidentialité. Pour répondre à ce verrou, nous utilisons une approche sémantique de représentation du bâtiment et particulièrement de la norme IFC. Cette approche permet d'extraire des fichiers IFC valides, partiels, à partir d'un fichier IFC modélisant l'intégralité du bâtiment (C7). Nous avons implémenté cette approche à travers un prototype dont nous avons évalué les performances (section 3.3.4.2.6). Ceci nous a permis de tirer des conclusions quant à la faisabilité d'extractions de données IFC avec des approches sémantiques.

La section suivante présente plus en détail les différents résultats obtenus par rapport aux contributions ci-dessus.

3.3.4.2 Résultats

3.3.4.2.1 Processus de conversion IFC EXPRESS vers OWL (C4)

Le Tableau 22 résume les règles implémentées par notre convertisseur pour la traduction de fichiers IFC 2x3 en OWL. Notre convertisseur exploite la version XML du standard IFC 2x3. L'ontologie ainsi générée est stockée dans un magasin de triples Stardog.

Nous ne nous sommes pas limités à la définition de nouvelles règles de conversion, mais avons comparé le résultat de notre conversion avec l'ontologie ifcOWL de bSI. Les deux ontologies sont générées à partir du schéma EXPRESS correspondant à la version 2x3 TC1 du standard IFC. Notre approche pour la construction de l'ontologie ifcOWL présente deux principaux avantages:

Les collections EXPRESS (e.g. constructions **`LIST`**) ne sont pas traduites en tant que **`RDFS List`** ou **`OWL List`**, comme c'est le cas pour la version officielle d'ifcOWL. Ces constructions sont traduites en tant que propriétés OWL, ce qui facilite l'accès aux données et améliore nettement le temps d'exécution des requêtes. Par exemple, l'attribut IFC **`coordinates`** est traduit, avec notre approche, sous la forme de trois propriétés OWL représentant les coordonnées sur les 3 axes respectivement **`coordinateX`**, **`coordinateY`** et **`coordinateZ`**. Notre approche permet ainsi d'éviter la composition de requêtes imbriquées (nécessaires pour parcourir les objets owl:List ou rdfs:List) pour récupérer la valeur de la coordonnée sur l'axe Z, par exemple.

<table>
<tr><td colspan="2">Entités IFC</td></tr>
<tr><td colspan="2">Description de la règle</td></tr>
<tr><td colspan="2">Une entité de base IFC est traduite en classe OWL. De cette manière, nous gardons la même hiérarchie d'entités que dans le modèle IFC.</td></tr>
<tr><td colspan="2">Attributs IFC</td></tr>
<tr><td>Description de la règle</td><td>Exemple</td></tr>
<tr><td>A partir de la séquence d'attributs dans une entité IFC, nous déterminons l'entité pointée par l'attribut. S'il s'agit d'un type défini (DEFINED TYPE), alors l'attribut est traduit en propriété données OWL. À partir des ensembles de départ et d'arrivée définis pour cet attribut, nous sommes capables de déterminer les domaines et portées de la propriété données OWL ainsi construite.</td><td>Le type défini <code>IfcAbsorbedDoseMeasure</code> devient la propriété données OWL <code>hasIfcAbsorbedDoseMeasure</code> qui prendra des valeurs de type <code>xsd:float</code>.</td></tr>
<tr><td>S'il s'agit d'une entité IFC, nous traduisons l'attribut IFC par une propriété objet OWL.</td><td>Le type <code>IfcContext</code> devient la classe OWL <code>IfcContext</code>, elle-même sousclasse de la classe OWL <code>IfcObjectDefinition</code> (qui correspond à la traduction de l'entité <code>IfcObjectDefinition</code>).</td></tr>
<tr><td colspan="2">Nous spécifions aussi les propriétés fonctionnelles (propriétés n'admettant qu'une seule valeur). Lorsque pour une relation IFC donnée, l'attribut <code>maxOccurs</code> vaut 1, alors la propriété objet OWL correspondante est identifiée comme étant fonctionnelle.</td></tr>
<tr><td colspan="2">Types SELECT</td></tr>
<tr><td>Description de la règle</td><td>Exemple</td></tr>
<tr><td>Dans le modèle IFC, le type SELECT est utilisé afin de permettre d'effectuer des sélections parmi des types de sélecteurs plus spécialisés. Dans ce cas, nous avons traduit le select type en tant qu'une classe OWL equivalente à l'union des types SELECT les constituant (sous-classes en OWL).</td><td>Le select type IfcUnit devient la classe OWL représentée par l'union des sous-classes OWL <code>IfcNamedUnit</code>, <code>IfcDerviedUnit</code> et <code>IfcMonetaryUnit</code>.
Le select type <code>IfcValue</code> devient la classe OWL représentée par l'union des sous-classes OWL <code>IfcSimpleValue</code>, <code>IfcMeasureValue</code> et <code>IfcDerviedMeasureValue</code>.</td></tr>
<tr><td colspan="2">Enumérations IFC</td></tr>
<tr><td>Description de la règle</td><td>Exemple</td></tr>
<tr><td>La norme IFC définit plusieurs énumérations, comme par exemple: <code>IfcCurrencyEnum</code>, <code>IfcSIPrefix</code>, etc.
Pour chacune de ces énumérations, nous les avons traduites en propriétés données OWL, dont la portée est définie à travers une liste contenant l'ensemble des valeurs possibles.</td><td>L'énumération <code>IfcSIPrefix</code> devient la propriété donnée <code>hasIfcSIPrefix</code> dont la portée est l'ensemble de valeurs suivant :
{<code>EXA^^xsd:string</code>, <code>PETA^^xsd:string</code>, <code>TERA^^xsd:string</code>, <code>GIGA^^xsd:string</code>, etc.}</td></tr>
</table>

Tableau 22. Règles implémentées par notre convertisseur pour la traduction de fichiers IFC 2x3 en OWL.

Nous avons choisi de ne pas traduire l'ensemble des IFC Defined Types en tant que classes OWL. Au lieu de cela, nous étudions les types de données contenus au sein des différents Defined Types et nous créons une seule classe OWL par type de données concerné. Par exemple, les types **`IfcVolumeMeasure`**, **`IfcAreaMeasure`**, **`IfcLengthMEasure`**, etc. sont tous traduits en tant que classe OWL appelée **`Real`** puisqu'ils contiennent des valeurs réelles. Dans notre modèle ifcOWL, cette classe **`Real`** possède une propriété de type de données, **`hasRealValue`** (ayant une portée dans **`xsd:double`**) permettant ainsi de spécifier le type de données concerné.

Lorsque nous comparons les caractéristiques de notre ontologie ifcOWL à l'ontologie ifcOWL de bSI (voir Tableau 23), on remarque un niveau d'expressivité plus bas, favorisant son intégration dans des applications, ainsi que la définition de propriétés de type de données fonctionnelles (50 avec notre approche contre 5 avec ifcOWL) de même que la définition de domaines et de portées pour de telles

propriétés types de données (respectivement 260 et 247 avec notre approche, contre seulement 5 et 5 avec ifcOWL). Le nombre de propriétés de types de données définies est effectivement bien plus important avec notre version d'ifcOWL qu'avec celle de bSI (247 contre 5).

Eléments de l'ontologie	**Notre ifcOWL** [CI18]	**ifcOWL bSI** (LDWG 2015)
Axiomes	7216	17783
Axiomes logiques	4455	11790
Classes	802	1093
Propriétés objet	1177	1422
Propriétés de type de données	247	5
Instances	0	1018
Expressivité DL	$\mathcal{ALUIF}^{(D)}$	$\mathcal{SHIQ}^{(D)}$
Axiomes de sous-classes	656	4344
Axiomes de classes équivalentes	46	2
Axiomes de classes disjointes	0	1888
Sous-propriétés objet	0	3
Propriétés inverses	111	96
Propriétés objet fonctionnelles	942	1292
Propriétés transitives	62	1
Domaines de propriétés objet	1073	1418
Portées de propriétés objet	1070	1418
Propriétés type de données fonctionnelles	50	5
Domaines de propriétés de type de données	260	5
Portées de propriétés de type de données	247	5
Axiomes d'assertions de classes	0	1313
Axiomes d'annotations d'assertions	639	2441

Tableau 23.Comparaison entre notre version de l'ontologie ifcOWL et celle utilisée par bSI, les deux étant basées sur le schéma EXPRESS IFC 2x3.

Nous avons montré qu'une telle conversion présentait des avantages en termes d'enrichissement sémantique et d'extraction partielle de sous-graphes [CI18]:

- Pour l'enrichissement sémantique, notre approche supporte la définition de nouveaux concepts par le biais de règles logiques (e.g. règles SWRL). Par exemple, si nous souhaitons spécifier un espace avec une ou plusieurs fenêtres, il suffit de définir la règle suivante :
  `IfcRelSpaceBoundary(?x) & IfcSpace(?y) & IfcWindow(?z) & RelatedBuildingElement(?x, ?z) & RelatingSpace(? x, ?y) ⇒ BimSpaceWithWindow(?y)`.
  Cette règle permet de définir un nouveau concept dans l'ontologie, BimSpaceWithWindow, concept qui va associer un concept IfcSpace et un concept IfcWindow. Une fois cette règle insérée dans l'ontologie, le mécanisme d'inférence va peupler l'élément BimSpaceWithWindow en calculant la liste des espaces avec fenêtres présents dans l'ensemble des données IFC contenues dans l'ontologie.
- La définition de tels concepts est une réponse au problème de l'hétérogénéité sémantique entre le vocabulaire de la norme IFC et le vocabulaire des besoins métiers. Dans de nombreux cas d'application, la norme IFC et les besoins métiers se trouvent en opposition. Le gestionnaire de patrimoine a besoin de manipuler des murs de façade, de vérifier s'il existe des pièces ou des espaces sans accès, ou de connaître le nombre d'espaces avec des fenêtres. À partir de la norme IFC, ces calculs sont souvent faits à l'aide d'algorithmes complexes qui se basent uniquement sur la géométrie des éléments contenus dans les IFC.
- Pour l'extraction de vues, notre approche exploite l'enrichissement sémantique avec les requêtes SPARQL. En effet, une fois que les différentes règles, définissant les nouveaux concepts, sont ajoutées à l'ontologie, les connaissances contenues sont automatiquement classifiées selon ces nouveaux concepts. Cela simplifie grandement l'écriture de requêtes SPARQL: au lieu de requêtes complexes (portant sur plusieurs concepts et relations), il est maintenant possible de requêter directement la base de connaissances pour obtenir, par exemple, l'ensemble des

instances d'espaces avec fenêtres (concept BimSpaceWithWindow défini à travers une règle SWRL).

3.3.4.2.2 Ontologie COBieOWL (C5)

C'est 2007 que l'approche COBie (*Construction-Operations Building information exchange*) a été proposée comme un moyen simple de partage d'informations bâtiment, informations qui n'incluent pas la modélisation géométrique. Il s'agit, en effet, d'un format de données à base de tableur (format ouvert) contenant des informations numériques d'un bâtiment. Créé par le National Institue of Building Sciences (NIBS) et plus particulièrement par le département en charge de la gestion de patrimoine (*Facility Maintenance and Operations Committee*), le projet COBie a été dirigé par un laboratoire de l'armée américaine, le Corps of Engineers. COBie permet l'échange d'informations entre plusieurs parties prenantes sans nécessiter de longues formations et de nouvelles acquisitions de logiciels. COBie a pour but d'améliorer la manière dont les informations sont capturées lors de la conception et de la construction de bâtiments, puis de les utiliser à des fins d'exploitation, de maintenance et de gestion des actifs. COBie est une norme reconnue dans le monde entier, implémentée dans des logiciels commerciaux de conception, de construction et de maintenance pour la gestion des informations de construction (East 2014). COBie offre une grande flexibilité car les informations sur le cycle de vie des bâtiments peuvent être visualisées dans des feuilles de calcul simples et peuvent être utilisées sur tous les projets de construction, indépendamment de leur taille et de leur complexité technologique. Un fichier COBie correspond à un bâtiment et structure les informations en plusieurs feuilles de calcul rassemblée au sein d'un même classeur (le fichier COBie). Un code couleur permet d'associer des contraintes aux différentes feuilles de calcul, par exemple la couleur jaune indique si la feuille de calcul est obligatoire, la couleur verte indique des feuilles spécifiées obligatoires, etc. Les contenus des colonnes d'une feuille de calcul COBie peuvent faire référence à des valeurs d'autres colonnes d'autres feuilles de calcul. Pour valider que les valeurs utilisées sont correctes, des règles de validation des données sont définies dans le modèle COBie. La figure suivante (Figure 41) illustre la définition d'une telle règle. Le modèle pour les feuilles de calcul COBie ainsi que des exemples de tels fichiers sont disponibles à partir de l'adresse suivante: http://www.nibs.org/?page=bsa_commonbimfiles

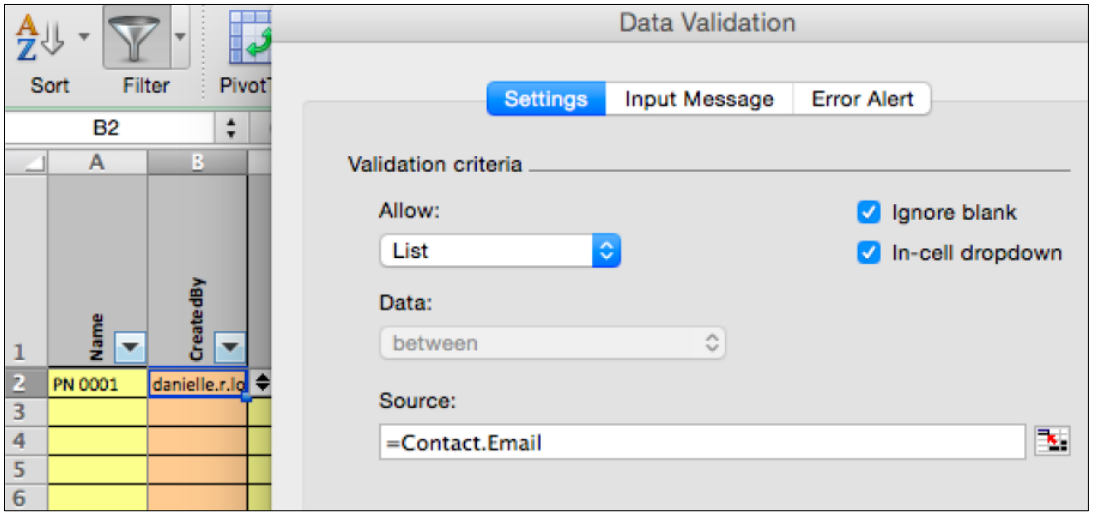


Figure 41. Illustration de la définition d'une règle de validation de données.

Avant de réfléchir à une adaptation de l'approche COBie en ontologie, nous avons étudié les différentes approches permettant de construire une ontologie OWL à partir de feuilles de calcul. En effet, considérant la structure du modèle COBie, il nous a semblé plus intuitif de générer une première ontologie à partir du modèle de feuilles de calcul, puis de manuellement augmenter l'ontologie avec les éventuelles classes, relations et contraintes manquantes. Les approches étudiées (O'Connor et al. 2010), (Jupp et al. 2010; Bowers et al. 2010) allaient toutefois à l'encontre de la logique derrière l'approche COBie (à savoir pas besoin d'apprendre un nouveau langage ou un nouveau logiciel). En effet, elles étaient toutes basées sur la définition de règles d'appariement entre les éléments des feuilles de calcul et ceux de l'ontologie. Chaque nouvelle version de COBie impliquerait donc la définition de nouvelles règles. De plus, les différents codes couleurs des feuilles de calcul n'étaient pas pris en compte par les approches étudiées.

Nous avons donc conçu une nouvelle approche permettant de construire l'ontologie COBieOWL de manière semi-automatique et en deux temps:

> Premièrement, nous avons défini des règles de conversion spécifiques, permettant de traduire chaque élément du modèle COBie 2.4 en tant qu'élément de l'ontologie. Ces règles sont détaillées dans notre publication [CI10], et ont été présentées à plusieurs sommets bSI. Grosso modo, chaque feuille de calcul devient une classe OWL et les cellules des différentes feuilles deviennent des valeurs de propriétés.
> Ensuite, le modèle ontologique ainsi construit (ou la TBox de l'ontologie COBieOWL) est peuplé avec instances et des données extraites du fichier COBie considéré. Nous avons construit un programme en Java permettant de peupler l'ontologie et d'ainsi construire son ABox. Pour ce faire, nous avons utilisé l'API Apache POI[22] (pour la gestion des feuilles de calcul) ainsi qu'OWL API[23] pour interagir avec l'ontologie.

La Figure 42 illustre une portion de l'ontologie COBieOWL ainsi construite. Pour chaque classe OWL `C`, correspondant à une feuille de calcul COBie, notre approche spécifie qu'elle est sous-classe de la classe **`CobieSheet`** (⊑ CobieSheet). Le concept **`CobieSheet`** est un concept que nous avons ajouté (CobieSheet ⊑ ⊤) et qui permet de représenter les concepts concrets (ou pouvant être instanciés) du standard COBie (e.g. **`Facility`**, **`Floor`**, **`System`**). Le Tableau 24 présente quelques métriques de cette ontologie.

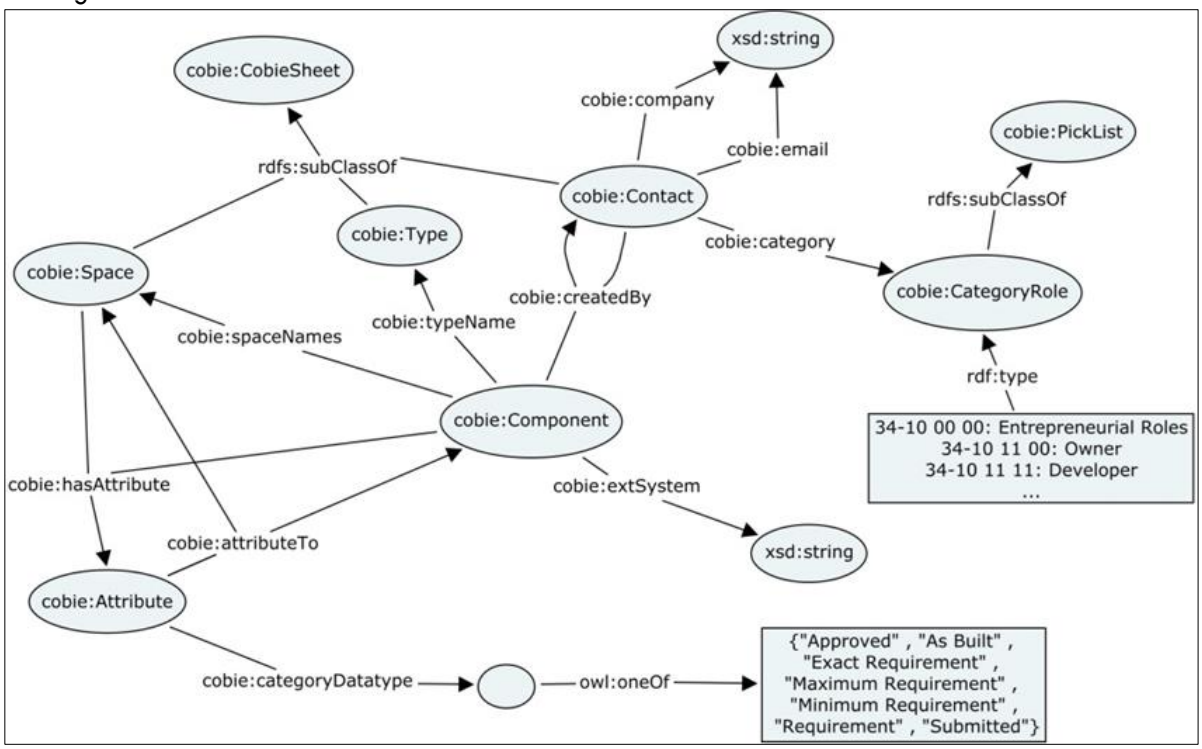


Figure 42. Illustation d'une portion de l'ontologie conçue pour le modèle COBie [CI10]

Par rapport à l'ontologie ainsi construite, plusieurs modifications et corrections ont été apportées directement en l'éditant avec Protégé. Pour ce faire, nous avons importé la sérialisation Turtle de l'ontologie dans Protégé, puis nous avons manuellement ajouté les différentes corrections nécessaires e.g. la portée de la propriété **`cobie:coordinateXAxis`** définie sur **`xsd:double`** et non **`xsd:string`** comme généré. La figure suivante donne un exemple d'une telle modification. Nous avons en effet dû préciser la portée de la propriété **`cobie:sheetName`** en tant que liste d'individus nommés extraits à partir des sous-classes de **`cobie:PickList`**.

[22] https://poi.apache.org/ 2
[23] http://owlapi.sourceforge.net/

Eléments de l'ontologie	COBieOWL [CI10]
Axiomes	19840
Axiomes logiques	19658
Classes	30
Propriétés objet	32
Propriétés de type de données	125
Propriétés inverses	7
Instances	9416
Axiomes de sous-classes	149
Sous-propriétés objet	25
Propriétés objet fonctionnelles	24
Domaines de propriétés objet	26
Portées de propriétés objet	24
Sous-propriétés type de données	123
Propriétés type de données fonctionnelles	122
Domaines de propriétés de type de données	124
Portées de propriétés de type de données	123
Triples dans la TBox	2503
Expressivité DL	$\mathcal{ALCHIF}^{(D)}$

Tableau 24. Métriques de l'ontologie COBieOWL conçue [CI10]

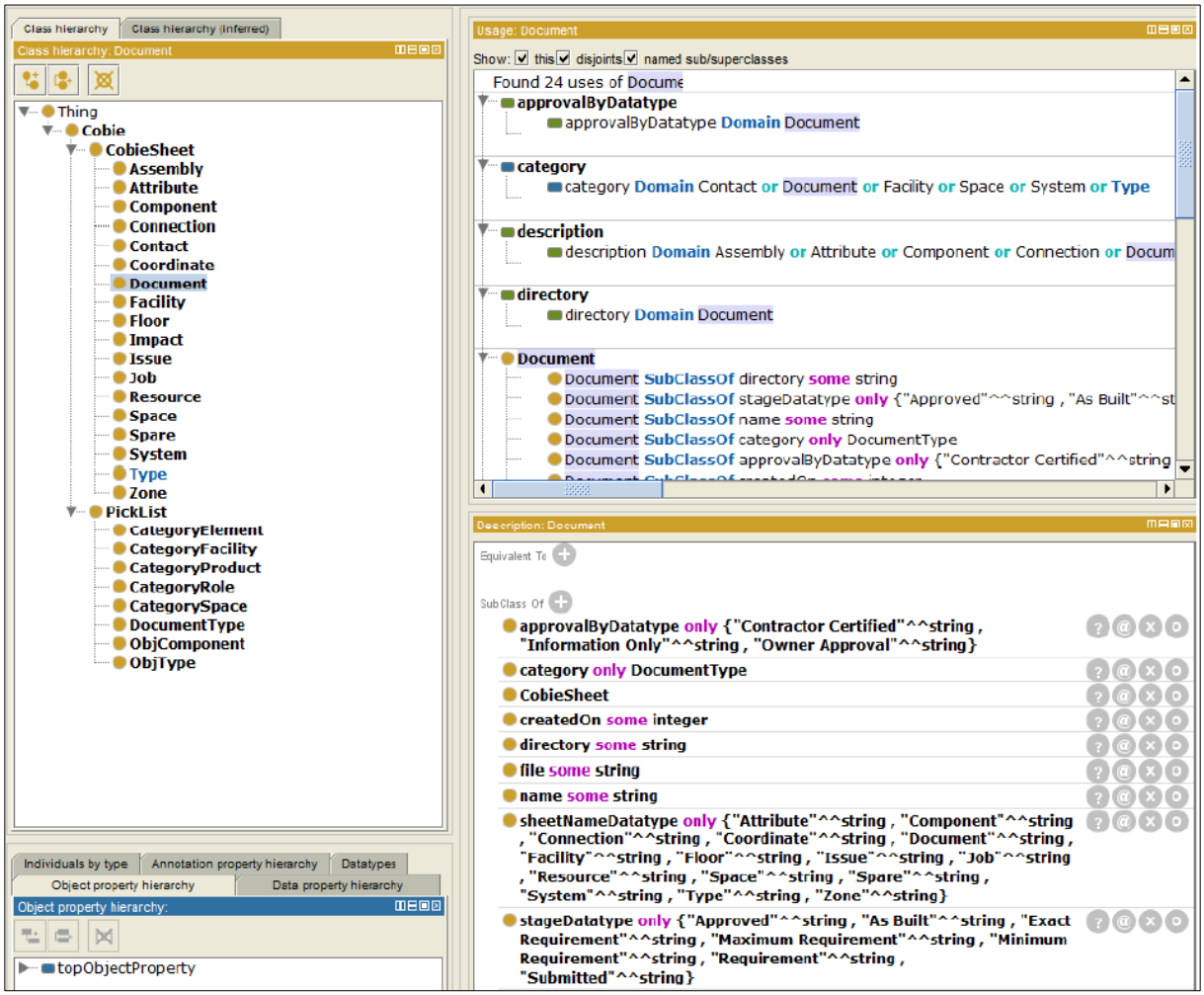


Figure 43. Modifications manuelles de l'ontologie COBieOWL dans Protégé [CI10]

Nous avons davantage étendu l'ontologie COBieOWL afin d'adresser le problème des feuilles de calcul avec des colonnes référençant des lignes dans d'autres feuilles de calcul. Pour adresser ce problème, nous avons suivi les étapes suivantes:

- Tout d'abord, nous avons identifié les noms de toutes les feuilles de calcul pouvant contenir des références à des lignes d'autres feuilles de calcul.
- Pour toutes les classes correspondantes (e.g. **`Document`**, **`Type`**), nous avons défini une nouvelle propriété objet OWL suivant le schéma de nommage **`[nomFeuille]To`**, avec pour:
  - Portée la classe **`cobie:CobieSheet`**, avec les restrictions suivantes (équation 14):

$$\begin{aligned} [nomFeuille] &\sqsubseteq \top \\ \top &\sqsubseteq \forall [nomFeuille]To.\,\text{CobieSheet} \\ \exists [nomFeuille]To.\top &\sqsubseteq [nomFeuille] \end{aligned}$$

Equation 14

  - Domaine la classe OWL correspondant à la feuille de calcul considérée

- Nous avons aussi ajouté la définition des propriétés inverses associées selon le schéma de nommage **`cobiehas[nomFeuille]`**

En termes d'avantages, cette approche ontologique du modèle COBie permet de:

- Requêter la base de connaissances ainsi construite, pour notamment extraire des sous-parties d'un fichier COBie. La requête suivante permet de récupérer toutes les informations bâtiment créées par un utilisateur ayant pour email "user1@email.com".

```
CONSTRUCT{
   ?x ?p ?o  }  WHERE {
   ?y a cobie:Contact.
   ?y cobie:email <mailto:user1@email.com>.
   ?x cobie:createdBy ?y.
   ?x ?p ?o }.
```

Figure 44. Exemple de requête SPARQL permettant d'afficher tous les éléments créés par un utilisateur donné

- Inférer de nouvelles informations, en appliquant un raisonneur au dessus de la base de connaissances. La figure suivante illustre les inférences faites sur la base de la relation entre les propriétés **`cobie:hasDocument`** et **`cobie:documentTo`** (qui sont l'inverse l'une de l'autre).

```
Assertions:
        cobie:documentTo(doc1,type1)
        cobie:hasDocument(type2, doc2)
Inferences:
        cobie:documentTo(type2,doc2)
        cobie:hasDocument(type1, doc1)
        (type1, type2 instances of cobie:Type)
        (doc1, doc2 instances of cobie:Document)
```

Figure 45. Inférences faites sur la base des caratéristiques de la propriété **`cobie:documentTo`** explicitement définies (a une propriété inverse **`cobie:hasDocument`**)

- Définir de nouveaux concepts et propriétés sur la base de ceux qui existent et en utilisant des règles logiques e.g. SWRL. La figure suivante illustre la définition du concept **`cobie:Window`**. La définition de tels concepts permet un appariement aisé vers des concepts synonymes dans d'autres ontologies (e.g. ifcWoD).
- Définir des liens vers des vocabulaires de données liées e.g. DBpedia et FOAF (équation 15).

$$\begin{aligned} cobie{:}\,Contact &\equiv foaf{:}\,Agent \\ \exists cobie{:}\,givenName.\top &\sqsubseteq foaf{:}\,Person \\ \exists cobie{:}\,familyName.\top &\sqsubseteq foaf{:}\,Person \\ cobie{:}\,givenName &\equiv foaf{:}\,givenName \\ cobie{:}\,familyName &\equiv foaf{:}\,familyName \\ cobie{:}\,email &\equiv foaf{:}\,mbox \end{aligned}$$

Equation 15

3.3.4.2.3 Ontologie ifcWoD (Web of Data) (C6)

De nos différentes études de la version ifcOWL choisie par bSI, nous avons pu déterminer que cette version n'exploite pas complétement les fonctionnalités OWL et continue d'être limitée par des constructions dictées par le langage EXPRESS (Tools 2015). Il faut en effet se rappeler que le standard IFC a d'abord été conçu dans un contexte de bases de données orientées-objet (Lee et al. 2012) et qu'un des principaux buts de la sérialisation STEP était de réduire la taille des fichiers échangés. C'est ce qui explique la syntaxe "compressée" de l'IFC: différentes propriétés d'une entité IFC sont un ensemble d'instances de la classe **`IfcPropertySingleValue`**, encapsulées au sein d'une instance de la classe **`IfcPropertySet`**. La Figure 46 illustre le contenu d'un tel fichier IFC STEP, où les différentes instances d'**`IfcPropertySingleValue`** (#2935, #2936, #2937 et #2941) sont regroupées au sein d'un **`IfcPropertySet`** (#2950). Cet ensemble de propriétés est associé à une instance de la classe **`IfcWallStandardCase`** (e.g. #3060) par le biais d'une instance **`IfcRelDefinesByProperties`** #14997.

```
#2935=IFCPROPERTYSINGLEVALUE('Retournement aux ouvertures',$, IFCINTEGER(0),$);
#2936=IFCPROPERTYSINGLEVALUE('Retournement aux extr\X\E9mit\X\E9s', $,IFCINTEGER(0),$);
#2937=IFCPROPERTYSINGLEVALUE('Largeur',$,IFCLENGTHMEASURE(0.32),$);
#2941=IFCPROPERTYSINGLEVALUE('Fonction',$,IFCINTEGER(0),$);
...
#2953=IFCMATERIAL('Mur par d\X\E9faut');
#2960=IFCMATERIALLAYER(#2953,0.32,$);
#2961=IFCMATERIALLAYERSET((#2960),'Mur de base:G\X\E9n\X\E9rique - 320 mm');
...
#2950=IFCPROPERTYSET('1Ks371pQHCOx_oJSZW78op',#34,'PSet_Revit_Type_Construction',
$,(#2935,#2936,#2937,#2941));
...
#3060=IFCWALLSTANDARDCASE('1iSKq$8HT2UvXyfHrxgRuh',#34,
'Mur de base:G\X\E9n\X\E9rique - 320 mm:193133',$,
'Mur de base:G\X\E9n\X\E9rique - 320 mm:168423',#3036,#3059,'193133');
...
#3085=IFCMATERIALLAYERSETUSAGE(#2961,.AXIS2.,.NEGATIVE.,0.16);
...
#14997=IFCRELDEFINESBYPROPERTIES('3NW$BjkfD1gh$ao1JHSikk',#34,$,$,
(#2890,#2906,#3002,#3060,#4605,#4685,#12594,#12656),#2950);
...
#14895=IFCRELASSOCIATESMATERIAL('0$aut9s4991QbgrydDAIH3',#34,
$,$,(#3060),#3085);
```

Figure 46. Une portion d'un fichier IFC STEP illustrant une relation **`IfcRelDefinesByProperties`** entre un objet IfcWallStandardCase et un ensemble de propriétés de type **`IfcPropertySingleValue`**.

Un tel référencement d'objets et de propriétés permet de fortement réduire la taille du fichier IFC STEP, mais ne s'applique que dans le contexte d'un seul fichier STEP.

Toutefois, lorsque l'on considère la conception d'une ontologie IFC, il convient d'utiliser correctement les classes et propriétés OWL, selon les bonnes pratiques de conception d'ontologies et sans forcémment chercher une correspondance totale avec le schéma EXPRESS. Lorsque l'on considère l'ontologie ifcOWL reconnue par bSI (LDWG 2015), sa structure de données présente plusieurs inconvénients:

- Elle ne facilite pas la compréhension humaine des différentes classes et propriétés IFC
- Elle complique l'application des principes des données liées
- Elle complexifie l'écriture des requêtes SPARQL

Sur la base de ces considérations, notre approche comprend trois principales modifications au niveau de la structure de l'ontologie IFC obtenue comme résultat:

- Les différentes relations définies par le standard IFC sont définies en tant que sous-classes d'IfcRelationship (41 sous-classes définies). Nous appliquons ainsi la définition de (Studer et al. 1998), considérant les ontologies en tant que spécifications formelles et explicites d'une conceptualisation partagée. Si l'on reprend l'exemple ci-dessous, notre approche consiste en la définition explicite

d'une relation (**ifcwod:isDefinedBy_ifcObject**) entre des instances de la classe **ifcowl:IfcWallStandardCase** et l'ensemble de propriétés considéré (instance de la classe **ifcowl:IfcPropertySet**). Ceci est illustré dans la Figure 47. Avec notre approche, nous considérons cette nouvelle propriété comme l'interprétation sémantique de l'entité IFC **IfcRelDefinesByProperties**.

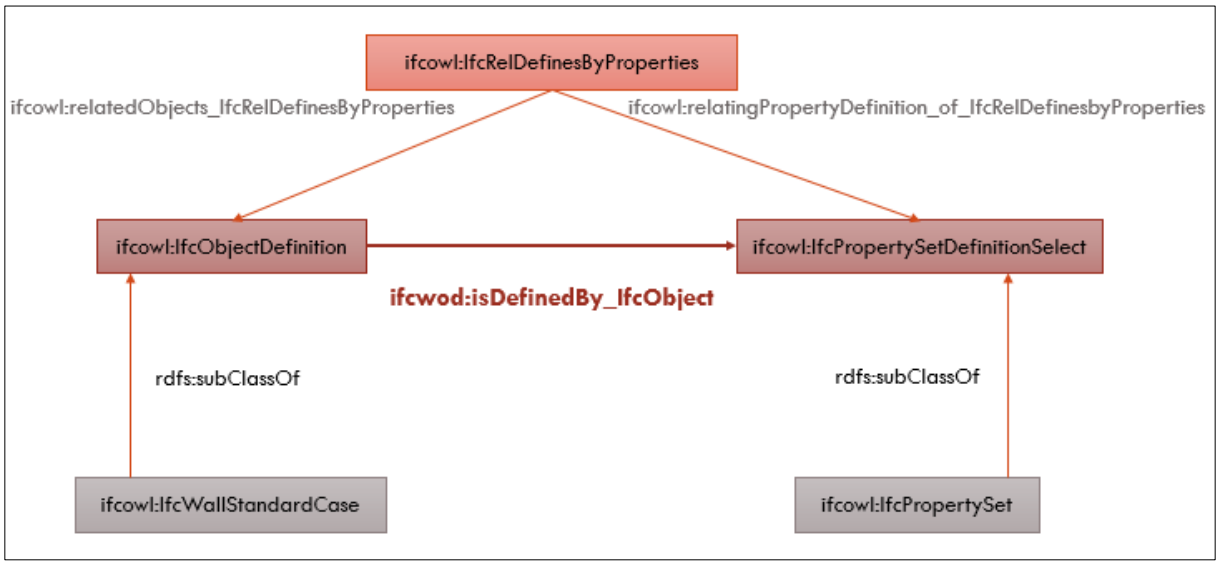


Figure 47. Définition de la relation isDefinedBy_IfcObject [CI11]

> L'entité **IfcProperty** est adaptée sémantiquement. Nous avons défini deux nouvelles sous-classes (**IfcSimpleProperty** et **IfcComplexProperty**) qui deviennent portée des propriétés objet OWL **ifcwod:hasSimpleProperty** et **ifcwod:hasComplexProperty**. Pour chaque sous-classe de **IfcSimpleProperty** et de **IfcComplexProperty**, les propriétés correspondantes sont définies en tant que sous-classe de **ifcwod:hasSimpleProperty** et respectivement **ifcwod:hasComplexProperty**. Le Tableau 25 résumé les différentes propriétés définies et leurs caractéristiques.

Propriétés ifcWoD	Domaine (rdfs:domain)	Portée (rdfs:range)	Sous-propriété de…
ifcwod:hasSimpleProperty	ifcowl:IfcPropertySet or ifcowl:IfcComplexProperty	ifcowl:IfcValue or ifcowl:ENUMERATION or ifcowl:IfcObjectReferenceSelect	owl:topObjectProperty
ifcwod:hasComplexProperty	ifcowl:IfcPropertySet or ifcowl:IfcComplexProperty	ifcowl:IfcComplexProperty	owl:topObjectProperty
ifcwod:hasReferenceValue	-	ifcowl:IfcObjectReferenceSelect	ifcwod:hasSimpleProperty
ifcwod:hasSingleValue	-	ifcowl:IfcValue	ifcwod:hasSimpleProperty
ifcwod:hasListValue	-	ifcowl:IfcValue	ifcwod:hasSimpleProperty
ifcwod:hasEnumeratedValue	-	ifcowl:ENUMERATION	ifcwod:hasSimpleProperty
ifcwod:hasTableValue	-	ifcowl:IfcValue	ifcwod:hasSimpleProperty
ifcwod:hasBoundedValue	-	ifcowl:IfcValue	ifcwod:hasSimpleProperty

Tableau 25. Propriétés ajoutées dans ifcWoD afin de représenter **IfcProperty** avec des propriétés objet OWL.

> La totalité des ensembles de propriétés (plus de 400 instances de type IfcPropertySet sont définies dans le standard) a été traduite en propriétés objet OWL. Pour ce faire, nous avons analysé leur définition XSD[24] et appliqué les règles listées dans le Tableau 26 pour obtenir les propriétés OWL associées.

[24] Disponible en ligne http://www.buildingsmart-tech.org/xml/psd/PSD_IFC4.xsd

Règle de correspondance	Définition XSD PropertySet	Adaptation en OWL
R1	`<xs:complexType name="PropertyDef">`	`owl:ObjectProperty`
R2	`<xs:element type="xs:string" name="Name">`	`rdfs:label` et URN
R3	`<xs:element type="xs:string" name="Definition">`	`rdfs:comment`
R4	`<xs:element name="NameAliases">`	`rdfs:label (@lang)`
R5	`<xs:element name="DefinitionAliases">`	`rdfs:comment (@lang)`
R6	`<xs:element type="PropertyType" name="PropertyType">`	`rdfs:subPropertyOf` et `rdfs:range`

Tableau 26. Propriétés définies dans IfcWoD afin d'adapter **`IfcProperty`** en OWL.

Lorsque comparé à l'approche ifcOWL choisie par bSI, notre approche présente plusieurs avantages, discutés ci-dessous.

Elle permet ***d'améliorer les capacités de raisonnement sur les données ainsi modélisées***, en tirant partie des constructions inhérentes à OWL. IfcWoD permet d'associer des caractéristiques logiques aux différentes propriétés OWL définies (e.g. **`owl:TransitiveProperty`**, **`owl:SymmetricProperty`**, **`owl:ReflexiveProperty`**), alors que l'approche ifcOWL ne considère que les inverses des propriétés OWL.

L'écriture et l'exécution de requêtes SPARQL sont grandement facilitées, principalement du fait des différentes sous-classes définies pour **`IfcRelationship`**. Comme illustré ci-dessous (Figure 48), lorsque l'on utilise le vocabulaire ifcWoD pour composer une requête SPARQL, celle-ci est simplifiée d'une part en utilisant la nouvelle propriété définie **`ifcwod:isPredecesorTo_IfcProcess`**, et d'autre part en exploitant le fait que cette propriété est définie comme transitive.

```
SELECT ?x ?z {
        ?x ifcowl:isPredecessorTo ?y.
        ?y ifcowl:relatingProcess_IfcRelSequence ?z. }

SELECT ?x ?y {
        ?x ifcwod:isPredecessorTo_IfcProcess ?y}
```

Figure 48. Exemple de simplification d'écriture de requêtes SPARQL.

L'adaptation sémantique des différentes instances de la classe **`IfcPropertyAbstraction`** présente quant à elle plusieurs avantages:

a) simplification additionnelle de l'écriture de requête,
b) amélioration du temps d'exécution des requêtes,
c) partage de propriétés dans un contexte données liées (à travers la définition explicite et formelle de la classe **`IfcProperty`**) et
d) réduction de la redondance des données.

Ces avantages peuvent être exemplifiés en considérant les 3 requêtes SPARQL ci-dessous (voir Figure 49). La colonne de gauche présente les requêtes exprimées uniquement en utilisant des termes du vocabulaire ifcowl, alors que la colonne de droite présente la syntaxe de ces mêmes requêtes utilisant des termes du vocabulaire ifcWoD et d'autres vocabulaires (ifcowl, RDF, etc.). L'exécution de ces requêtes a été évaluée dans un environnement technique constitué d'un serveur (Intel Xeon E5-2430 à 2.2GHz et 8Gb de mémoire RAM DDR3), avec une instance d'une base de connaissances Stardog exploitant 2 des 6 cœurs du processeur, et une machine client (Intel Core i7-4790 à 3.6GHz et 8Gb de mémoire RAM DDR3). Dans cet environnement, chaque requête a été exécutée 30 fois par la machine client et sur la base de connaissances correspondant au bâtiment de l'entreprise partenaire (fichier IFC-STEP d'une taille de 11Mb, traduit en ifcowl et en ifcWoD). Le Tableau 27 comprend les temps d'exécution moyens ainsi que leur écarts-types, pour chacune des deux approches. En utilisant notre approche ifcWoD, on réduit le temps d'exécution respectivement de 89,26%, de 95,15% et de 95,85% pour les 3 requêtes. On note aussi que cette réduction du temps d'exécution ne se fait pas au détriment des résultats retournés.

La réduction de la redondance des données peut être démontrée en considérant la propriété **`isExternal`**, telle que définie dans l'ensemble des propriétés IFC pour les murs (**`Pset_Wall-Common`**). La Figure 50 présente comment cette propriété est modélisée avec l'approche ifcowl et comment elle l'est avec notre approche ifcWoD. Dans le premier cas de figure, plusieurs assertions sont répliquées dans la base de connaissances (e.g. **`:PropertyInstance ifcowl:name_IfcProperty "IsExternal"^^xsd:string`**) et ce pour chaque instance de la classe **`IfcPropertySet`** devant contenir une propriété "**`isExternal`**". Ceci est évité avec notre approche.

Querying solely with ifcOWL terms

Q1:
```
SELECT ?wall WHERE {
?wall rdf:type ifcowl:IfcWall;
      ifcowl:isDefinedBy_IfcObject ?rel.
?rel  ifcowl:relatingPropertyDefinition
    -_IfcRelDefinesByProperties ?pSet.
?pSet ifcowl:hasProperties_IfcPropertySet ?p.
?p    rdf:type ifcowl:IfcPropertySingleValue;
      ifcowl:name_IfcProperty ?name.
?name expr:hasString "IsExternal"^^xsd:string.
?p ifcowl:nominalValue_IfcPropertySingleValue ?val.
?val  ifcowl:hasBoolean "true"^^xsd:boolean. }
```

Q2:
```
SELECT ?door ?reference WHERE {
?door rdf:type ifcowl:IfcDoor;
      ifcowl:isDefinedBy_IfcObject ?rel.
?rel  ifcowl:relatingPropertyDefinition ?pSet.
?pSet ifcowl:hasProperties_IfcPropertySet ?p.
?p    rdf:type ifcowl:IfcPropertySingleValue;
      ifcowl:name_IfcProperty ?name.
?name expr:hasString "Reference"^^xsd:string.
?p    ifcowl:nominalValue_IfcPropertySingleValue ?val.
?val   expr:hasString ?reference. }
```

Q3:
```
SELECT ?x ?reference WHERE {
?floor rdf:type ifcowl:IfcBuildingStorey.
?floor ifcowl:elevation_IfcBuildingStorey ?elev.
?elev  expr:hasDouble ?y.
?floor ifcowl:isDecomposedBy_IfcObjectDefinition ?rel.
?rel   ifcowl:relatedObjects_IfcRelAggregates ?x.
?x     rdf:type ifcowl:IfcSpace;
       ifcowl:isDefinedBy_IfcObject ?rel.
?rel   ifcowl:relatingPropertyDefinition
        -_IfcRelDefinesByProperties ?pSet.
?pSet  ifcowl:hasProperties_IfcPropertySet ?p.
?p     rdf:type ifcowl:IfcPropertySingleValue;
      ifcowl:name_IfcProperty ?name.
?name  expr:hasString "Reference"^^xsd:string.
?p     ifcowl:nominalValue_IfcPropertySingleValue ?val
?val   expr:hasString ?reference. FILTER (?y > 0) }
```

Querying with IfcWoD terms

Q1':
```
SELECT ?wall WHERE {
?wall rdf:type ifcowl:IfcWall;
      ifcwod:isDefinedBy_IfcObject ?pSet.
?pSet pset_WallCommon:isExternal ?val.
?val  expr:hasBoolean "true"^^xsd:boolean. }
```

Q2':
```
SELECT ?door ?reference WHERE {
?door rdf:type ifcowl:IfcDoor;
      ifcwod:isDefinedBy_IfcObject ?pSet.
?pSet pset_DoorCommon:reference ?val.
?val  expr:hasString ?reference. }
```

Q3':
```
SELECT ?x ?reference WHERE {
?fl   rdf:type ifcowl:IfcBuidingStorey.
?fl   ifcowl:elevation_IfcBuildingStorey ?elev.
?elev expr:hasDouble ?y.
?fl ifcwod:isDecomposedBy_IfcObjectDefinition ?x.
?x    rdf:type ifcowl:IfcSpace;
      ifcwod:IsDefinedBy_IfcObject ?pSet.
?pSet pset_SpaceCommon:reference ?val.
?val  expr:hasString ?reference. FILTER (?y > 0)}
```

Figure 49. Comparaison entre l'écriture de requête avec ifcOWL (à gauche) et IfcWoD (à droite).

	Q1	**Q1'**	Q2	**Q2'**	Q3	**Q3'**
Moyenne (en s)	0,242	**0,026**	0,516	**0,025**	1,348	**0,056**
Ecart-type	0,024	**0,009**	0,019	**0,008**	0,024	**0,017**
Nombre de résultats	37	**37**	141	**141**	67	**67**
% de réduction moyen	-	**89,26%**	-	**95,15%**	-	**95,85%**

Tableau 27. Analyse des performances associées à l'exécution de requêtes.

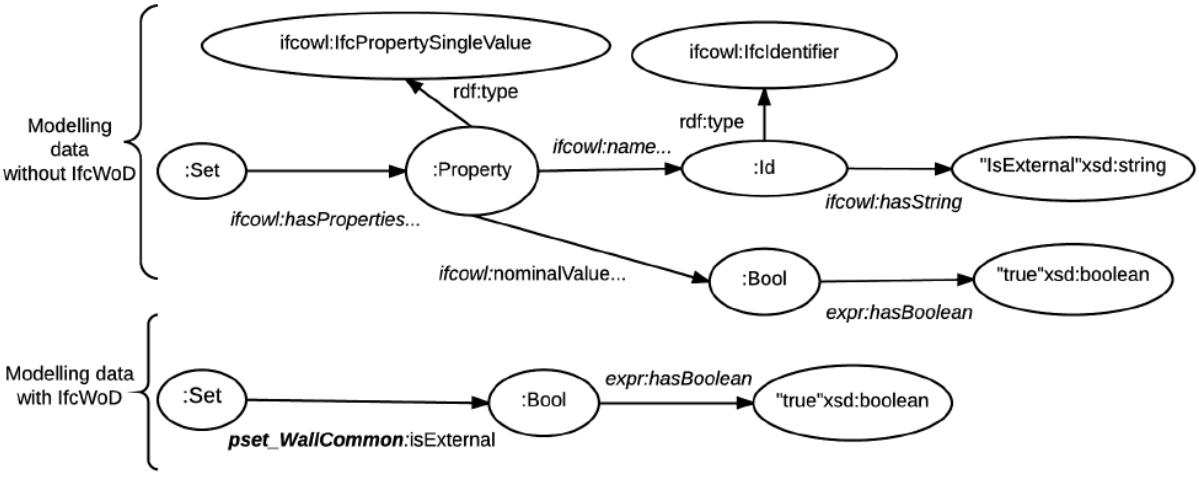


Figure 50. Comparaison des modélisations de la propriété **`IsExternal`** du **`Pset_WallCommon`** – en haut sans IfcWoD, en bas, avec IfcWoD.

Nous approche ifcWoD se focalise sur la modélisation correcte des relations IFC dans une ontologie OWL. Pour ce faire, nous avons basé notre proposition sur des bonnes pratiques et des recommandations largement reconnues et acceptées dans le domaine de la conception d'ontologies. Notre ontologie IfcWoD se base sur une nouvelle méthode semi-automatique de conception d'ontologies, qui est plus appropriée pour lier sémantiquement des données IFC dans un contexte Web de données (ou données liées). Les choix de modélisation appliqués dans l'ontologie ifcWoD se basent sur une étude rigoureuse des choix implémentés par les approches identifiées dans l'état de l'art. Sur la base des inconvénients identifiés, nous proposons une nouvelle modélisation permettant une application simplifiée des principes Linked (Open) Data. Il ne s'agit pas d'une autre version d'ifcOWL, mais d'une nouvelle ontologie utilisant des termes du vocabulaire ifcOWL. En outre, un sous-ensemble d'ifcOWL peut être considéré en tant que méta-modèle de notre ontologie IfcWoD.

3.3.4.2.4 Architecture fédérée faiblement couplée pour l'interopérabilité entre ontologies OWL - FOWLA (C1)

Avec le développement des technologies du Web sémantique et le nombre croissant de leurs applications, l'hétérogénéité et le nombre d'ontologies existantes augmentent de manière exponentielle. Cela soulève le problème de comment déterminer si deux ontologies ont la même signification (Euzenat & Shvaiko 2013; Shvaiko & Euzenat 2013). Répondre à cette question revient à effectuer une mise en correspondance d'ontologie ou *ontology matching*. Un tel processus revient à identifier des correspondances entre les éléments de différentes ontologies (Euzenat & Shvaiko 2013). Un ensemble de telles correspondances entre deux ou plusieurs ontologies définit un alignement d'ontologies. L'alignement est le résultat du processus d'ontology matching (Euzenat & Shvaiko 2013).

Lors d'une mise en correspondance d'ontologies, la complexité du processus est impactée par:
- La prise en compte des conséquences de chacune des ontologies impliquées.
  - Par conséquence d'une ontologie il faut entendre l'ensemble des assertions implicitement générées par une ontologie (Euzenat 2011). Ces assertions impactent directement l'exécution de requêtes au-dessus de l'ontologie considérée.
- Le fait que les langages OWL (Dean et al. 2006) et OWL 2 (Motik et al. 2012) ne fournissent pas des représentations suffisamment expressives des connaissances nécessaires à la définition d'alignements corrects.
  - Pour illustrer ceci, prenons l'exemple du langage OWL-DL: celui-ci ne permet pas de définir la propriété oncleDe à partir des propriétés frereDe et parentDe. En effet, les règles axiomatiques ne sont pas supportées par OWL-DL seul (Bruce 2011). Nous pouvons par contre combiner des représentations en OWL-DL avec des règles SWRL et ainsi déclarer des clauses de Horn arbitraires au-dessus d'ontologies OWL-DL. Néanmoins, cela peut conduire à une indécidabilité dans des tâches de raisonnement cruciales, telles que la subsomption en OWL (Baroglio et al. 2008). Afin de retrouver la décidabilité dans de tels contextes, les restrictions DL-safe (Boris Motik et al. 2005; Motik 2006) sont mises en œuvre.

Dans ce contexte, et dans ce qui suit, l'interopérabilité entre ontologies sera interprétée comme la capacité de partager des données entre différentes ontologies. Comme mentionné précédemment, l'interopérabilité des ontologies peut être réalisée grâce à des alignements d'ontologies, exprimés à travers différents formats d'alignement existants. L'un des avantages des formats d'alignement d'ontologies est qu'ils sont exprimés avec des langages plus faciles à manipuler et plus expressifs que les langages basés sur OWL. Les règles SWRL peuvent être considérées comme un format d'alignement: elles spécifient les alignements d'ontologies au moyen de clauses de Horn positives (comportent un littéral positif et aucun littéral négatif). Les règles SWRL composant un alignement d'ontologies ne sont pas des règles de ré-écriture, et elles ne génèrent donc pas des données redondantes (Euzenat & Shvaiko 2013).

Afin de mieux comprendre notre approche pour l'interopérabilité entre ontologies, prenons l'exemple de la Figure 51 : deux ontologies (Onto1 et Onto2) sont appariées à travers un alignement exprimé avec un ensemble de règles SWRL (*rule set*), au-dessus des deux ontologies. Notre approche se veut un compromis entre l'optimisation du temps d'exécution de requêtes et la réalisation des tâches d'inférence sur les ontologies considérées. Avec notre approche, il est possible de sélectionner uniquement les alignements pertinents par rapport à la requête à exécuter, tout en limitant l'inférence à la déduction des seules données nécessaires à la requête.

Alors que les différentes approches étudiées, (Makris 2010) ou (Correndo et al. 2010), appliquent des procédures de ré-écriture de la requête afin qu'elle ne contienne que des termes d'une ontologie, notre approche permet d'exécuter des requêtes comprenant des termes issus des deux ontologies. En ce sens, il importe peu, dans notre cas, de définir une ontologie source et une ontologie de destination. Dans notre approche, ***l'ontologie source est déterminée à partir de là où est adressée la requête***: si la requête est adressée à Onto1, alors Onto1 devient l'ontologie de destination et Onto2 est considérée ontologie source. Selon là où est adressée la requête, il est possible qu'une ontologie soit à la fois source et destination (c'est notamment le cas si la requête $Q = \{x \mid x \in C \ \wedge x \ \in A\}$ est adressée aux ontologies de la Figure 51).

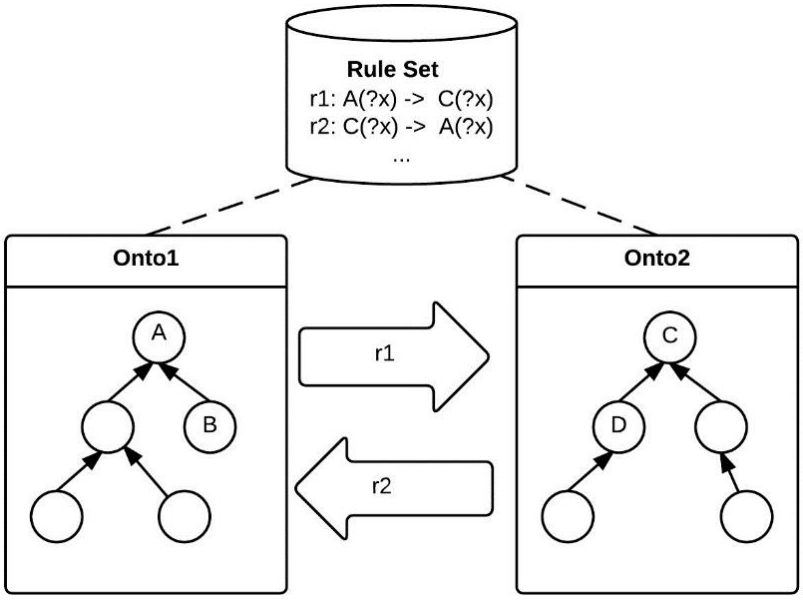


Figure 51. Exemple d'interopérabilité entre ontologies implémentée par le biais de règles.

Nous avons formalisé notre approche afin de prendre en compte plus que deux ontologies et afin de pouvoir fédérer tout ensemble d'ontologies alignées par le biais de règles SWRL. La Figure 52 illustre la vue d'ensemble.

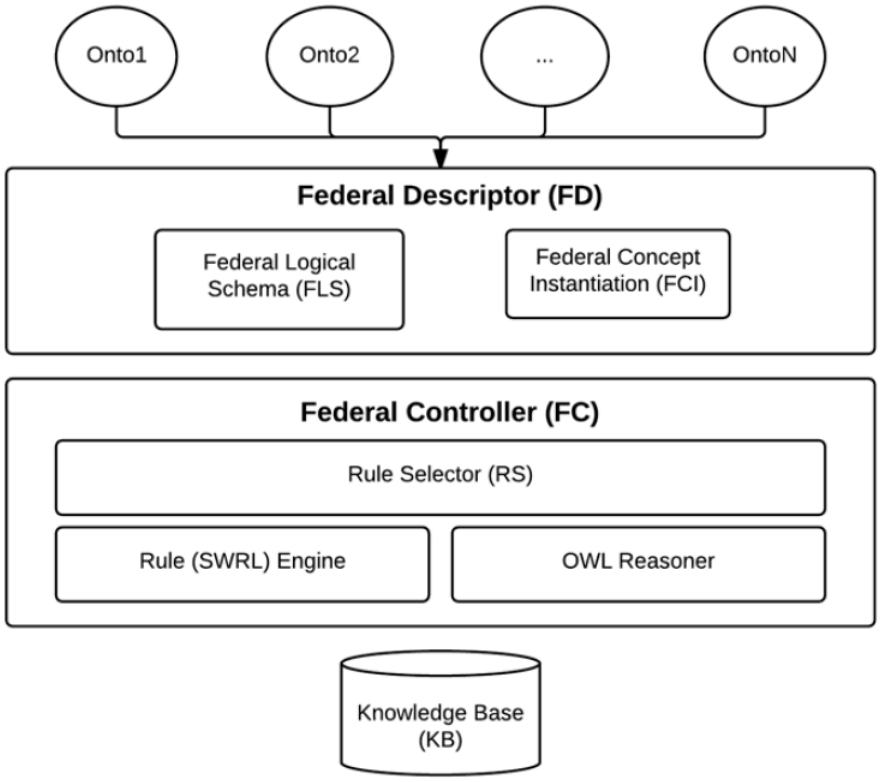


Figure 52. FOWLA vue d'ensemble.

Notre architecture pour la fédération d'ontologies faiblement couplée (ou FOWLA) comprend deux principaux composants:

- Le module FD (*Federal Descriptor*) est en charge des alignements entre les différentes ontologies considérées. Ce module comprend deux sous-modules:
  - Le composant FLS (*Federal Logical Schema*) comprend les ensembles de règles logiques décrivant les alignements entre ontologies. S'agissant de règles SWRL, elles ne permettent pas l'instanciation de nouveaux concepts dans les ontologies respectives (à cause des problèmes d'indécidabilité lors de l'intégration OWL avec SWRL, nous utilisons des règles DL-safe).
  - Le composant FCI *(Federal Concept Instantiation*) vient palier à ce manquement. Les règles DL-safe ne permettant de manipuler que des instances déjà présentes dans la base de connaissances, le module FCI crée l'ensemble des instances nécessaires afin que l'ontologie de destination puisse encapsuler des données de l'ontologie source. Pour ce faire, le module FCI instancie les classes nécessaires et ajoute les assertions de propriétés nécessaires à la base de connaissances (obtenue en unifiant les TBox des différentes ontologies)
- Le module FC (*Federal Controller*) est exécuté au moment de la requête et permet d'échanger des données entre les ontologies, selon l'alignement généré par le module FD. Toujours au moment de la requête, ce composant se charge aussi de vérifier les politiques d'accès associées aux données. Le module FC effectue les inférences nécessaires pour répondre à une requête impliquant des éléments d'une ou plusieurs ontologies fédérées. Pour ce faire, le module FC comprend un sélecteur de règles ou RS (Rule Selector). Le composant RS a été spécifiquement conçu pour les raisonneurs à chaînage arrière, et permet de sélectionner uniquement les règles SWRL nécessaire pour l'exécution de la requête considérée. Son fonctionnement est détaillé dans la section suivante.

Lorsqu'une requête SPARQL est adréssée, notre architecture FOWLA la traite en deux temps: une première phase de prétraitement (pour l'exécution des algorithmes du module FD), puis une deuxième phase d'exécution de la requête (exploitant le module FC pour retourner les données pertinentes à partir des ontologies visées). Ces deux phases sont illustrées dans la Figure 53.

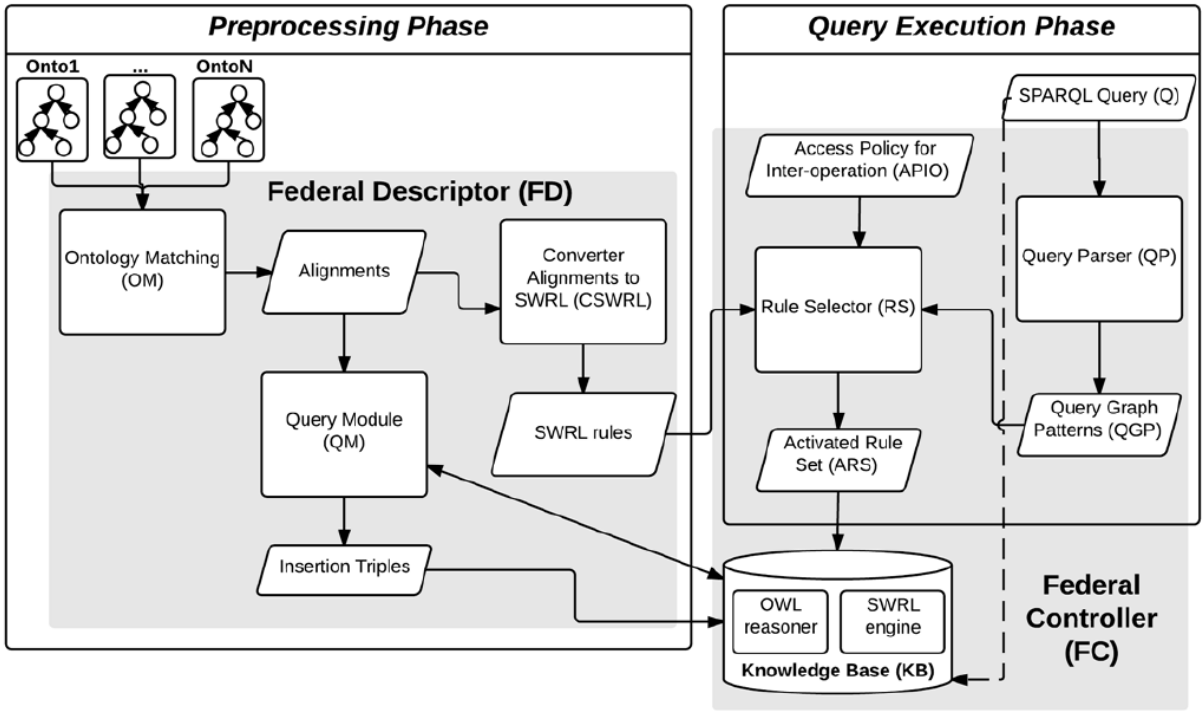


Figure 53. FOWLA composants des phases de prétraitement et d'exécution de requêtes.

Lors de la phase de pré-traitement, et selon le nombre d'ontologies considérées, il peut être nécessaire de construire plusieurs alignements complexes. La mise en correspondance d'ontologies est en effet un processus connu pour être complexe et fastidieux. Nous recommandons l'utilisation au préalable d'outils tels ASMOV (Shvaiko & Euzenat 2013) pour aider dans la définition de tels alignements. Il faut garder à l'esprit que la qualité des résultats des approches automatiques pour la mise en correspondance

d'ontologies dépend grandement du niveau d'implication de l'utilisateur dans la vérification et la validation des alignements produits. De plus, comme ces solutions automatiques ne prennent pas en charge la génération d'alignements complexes (e.g. un sous-graphe d'Onto1 mappé en tant que propriété dans Onto2), l'implication de l'utilisateur dans la définition d'alignements demeure cruciale (Shvaiko & Euzenat 2013).

Une fois définies les règles formant les différents alignements entre ontologies, si le format d'alignement n'est pas basé sur des règles (e.g. SWRL), un processus de conversion est exécuté (module de conversion CSWRL dans la Figure 53). Les alignements résultants sont exprimés à travers des règles SWRL et sont inclus dans le sous-module FLS (représenté par la case "SWRL rules" sur la Figure 53).

Par la suite, le module de requête (QM) identifie chaque alignement présentant une hétérogénéité de schéma et nécessitant donc des instanciations de classes et des assertions de propriétés pour modéliser des données provenant d'autres ontologies. Le module QM récupère les instances sans assertions de propriétés pour relier les données d'une ontologie source vers une ontologie cible. Pour ce faire, il s'appuie sur des requêtes SPARQL adressées à la base de connaissances (KB pour *Knowledge Base* dans la Figure 53). Nous avons choisi d'utiliser des requêtes SPARQL car la solution choisie pour héberger nos ontologies est la base de données graphe Stardog qui ne supporte que le langage de requête SPARQL. Le module QM génére uniquement les assertions nécessaires à la résolution d'hétérogénéités de schémas; les triples associés sont effacés de la base de connaissances lorsque le contenu du module FLS change c'est-à-dire lorsque les alignements entre ontologies sont modifiés.

Une fois les tâches du module FD effectuées, nous sélectionnons le sous-ensemble de règles nécessaire pour l'exécution de la requête considérée. Pour ce faire, nous avons développé un analyseur de requête SPARQL ou QP (*Query Parser*). Ce composant reçoit la requête SPARQL considérée (voir Figure 53) et identifie les différents éléments qu'elle contient e.g. classes et propriétés OWL. En exploitant les restrictions de domaine et de portée des différentes propriétés identifiées, le module RS (Rule Selector - détaillé dans la section suivante) sélectionne les règles SWRL nécessaires à l'exécution de la requête. Tout d'abord les règles du module FLS sont filtrées et seules sont sélectionnées les règles permettant d'inférer des triples pour les classes et propriétés de la requête (sélection effectuée par le module Query Graph Pattern ou QGP dans la Figure 53). Ensuite, le module RS identifie les règles ayant la même propriété dans leur antécédent, et sélectionne uniquement celles qui respectent les contraintes de domaine et de portée définies dans la requête. L'ensemble de ces règles forme l'ARS (Activated Rule Set) et représente l'unique ensemble de règles pris en compte lors de l'exécution de la requête SPARQL considérée. Les différentes phases de notre processus d'optimisation de l'exécution de telles requêtes SPARQL sont détaillées dans la section suivante (voir la 0 présentant notre algorithme de sélection de règles).

Dans ce qui suit, je vais brièvement discuter les avantages associés à l'implémentation de cette approche. Afin de mieux les comprendre, considérons le cas de figure suivant (voir Figure 54), où 4 ontologies (A, B, C et D) sont alignées par le biais de règles SWRL (FD(A,B) représentant l'alignement entre les ontologies A et B).

> FOWLA permet de ***déduire de nouveaux alignements entre ontologies*** - Pour illustrer cet avantage, supposons les ontologies et leurs alignements tels qu'illustrés dans la Figure 54. Pour chacune de ces ontologies, soit IS(X,Y) le sous-graphe de l'ontologie X nécessaire pour interopérer avec l'ontologie Y (IS pour *Interoperable Schema*). En supportant l'inférence d'assertions d'une ontologie à l'autre, nous réduisons le nombre d'alignements (ou règles) à définir. Si l'on considère l'ontologie C comme entièrement peuplée et le sous-graphe IS(C,B) comme équivalent de l'ontologie C, alors, une fois définies les règles composant FD(B,C), il est possible d'accéder aux assertions de C en requêtant uniquement B. En outre, les règles composant FD(A,B) permettent aussi d'accéder aux assertions de C en requêtant l'ontologie A. L'ontologie A devient dans ce cas l'intersection des deux sous-graphes IS(B,A) et IS(B,C). Il n'est plus nécessaire de définir FD(C,A) puisque les ontologies A et C sont alignées indirectement et totalement intégrées via le composant FC qui peut exploiter les règles présentes dans FD(C,A) et dans FD(B,C) pour inférer l'alignement entre A et C.

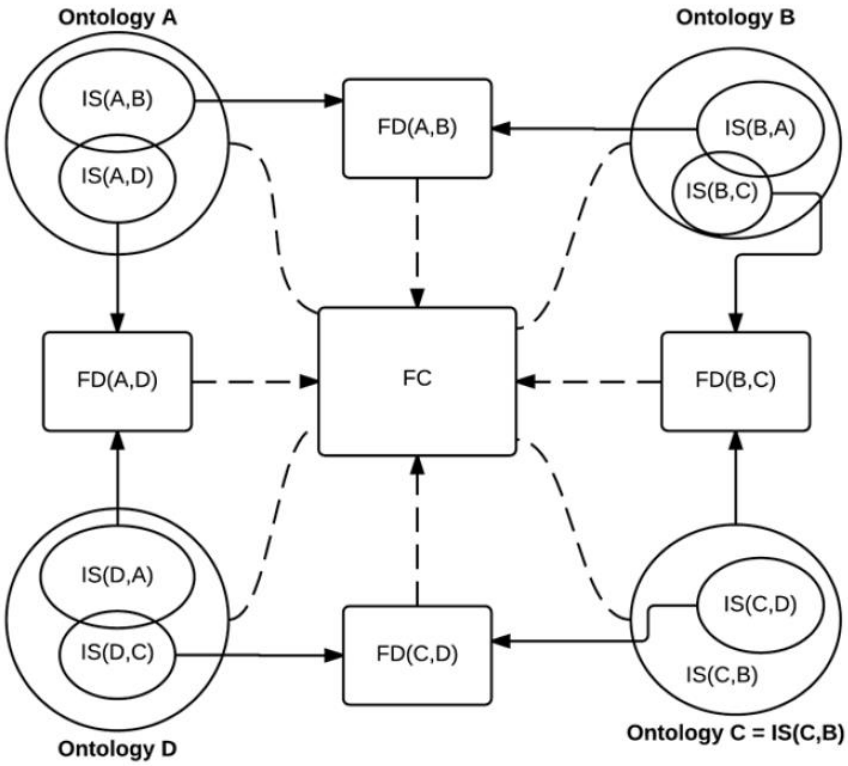


Figure 54. Exemple d'application de FOWLA pour l'interopérabilité de quatre ontologies A, B, C et D.

- FOWLA ***évite la redondance des données*** - Notre composant FC est basé sur un raisonnement employant le chaînage arrière (Russell & Norvig 2009). Ainsi les ontologies fédérées ne répliquent pas les données d'interopérabilité. Ceci s'explique à travers le fait qu'un raisonneur basé sur le chaînage arrière effectue les tâches d'inférence lors de l'exécution de la requête, sans avoir besoin de matérialiser ces données. Les données modélisées avec une ontologie donnée sont rendues disponibles aux autres ontologies via l'inférence ou en appliquant les alignements à base de règles via un raisonneur chaînage arrière. De plus, lorsque les assertions d'une ontologie source sont modifiées, le composant FC infère les triples impactés par la modification dans l'ontologie cible lors de la prochaine exécution de requête.
- FOWLA supporte une ***maintenance modulaire***, en préservant l'autonomie entre les différentes ontologies - Supposons une modification au niveau du schéma (TBox) de l'ontologie A, plus particulièrement au niveau du sous-graphe IS(A,D) - IS(A,B) intersecté avec IS(A,D). Dans ce cas, les seuls éléments devant évoluer sont FD(A,D) et IS(D,A). Ainsi une interoperabilité totale est préservée entre A, D et les autres ontologies. Les autres composants FD et ontologies demeurent inchangés. Seul le système basé sur l'ontologie A doit évoluer, les systèmes utilisant les ontologies B, C et D ne sont pas impactés, leur autonomie est donc préservée.
- FOWLA supporte ***l'exécution de requêtes composées avec des termes issus de l'ensemble des ontologies*** alignées
- FOWLA permet ***d'améliorer le temps d'exécution des requêtes SPARQL***

Pour l'illustration des deux derniers avantages listés ci-dessus, le lecteur est prié de consulter la section 3.3.4.2.6 Evaluation. La section suivante détaille le fonctionnement du module de sélection de règles et discute ses avantages pour l'optimisation de l'exécution de requêtes SPARQL au dessus d'alignements d'ontologies exprimés par le biais de règles SWRL.

3.3.4.2.5 Algorithme pour la sélection de règles logiques (C2)

Exprimer des alignements d'ontologies à l'aide des règles de Horn (axiomes de règles) (Euzenat & Shvaiko 2013) permet de résoudre le problème d'hétérogénéité. En effet, plusieurs recherches ont identifié la combinaison de règles axiomatiques et de logiques de description en tant que solution pour parvenir à l'interopérabilité parmi les ontologies (Sowa 1999). Lors de l'adressage d'une requête à un alignement d'ontologie, les clauses de Horn doivent être interprétées. Cela se traduit par la matérialisation des faits dérivés issus de ces règles. Selon le type de raisonneur utilisé, les cas de figure suivants existent:

- Dans le cas de raisonneurs basés sur du chaînage avant, l'ensemble des faits impliqués par les différentes clauses de Horn est pré-calculé, ce qui augmente considérablement la taille des données administrées.
- Les raisonneurs basés sur du chaînage arrière (Russell & Norvig 2009; Gallier 2015) interprètent les différentes clauses de Horn au moment de l'exécution de la requête (pas de pré-calcul). Pour ce faire, ils utilisent une méthode de résolution appelée "SLD" pour "*Selective Linear Definite*" (Apt 1997). Cette méthode permet d'interpréter la requête en tant que négation de conjonction de sous-requêtes.

L'interprétation des règles SWRL est aussi différente selon les types de raisonneurs et, d'après notre état de l'art, il existe quatre principales approches présentées à la section 3.1.4.3.3 (Introduction).

Déduire des faits implicites à partir de règles SWRL et ce lors de l'exécution de requêtes SPARQL peut augmenter considérablement le temps de traitement d'une telle requête. De plus, l'ensemble des raisonneurs considérés pour effectuer des inférences lors de l'exécution de requêtes sont des raisonneurs à usage général OWL/SWRL (voir Tableau 10) c'est-à-dire ils ne sont pas optimisés ou conçus pour l'alignement d'ontologies. Ces raisonneurs ne font pas la distinction entre les règles utilisées pour l'alignement entre ontologies et les règles inhérentes à tout langage d'ontologie et utilisées pour l'inférence (indépendamment que l'ontologie soit alignée ou pas).

Afin d'optimiser le temps d'exécution des requêtes adressées sur des ontologies alignées par le biais de règles SWRL, notre idée à été de spécifier une ***procédure pour la ré-écriture des règles SWRL***. C'est le module de Pré-Traitement des règles qui est en charge de leur ré-écriture. Un raisonneur généraliste pourra alors traiter uniquement les règles pertinentes pour répondre à une requête donnée. Ainsi, nous réduisons le nombre de règles considérées lors de l'exécution des requêtes. Les différents tests réalisés ont prouvé que cela permet de réduire le temps d'exécution des requêtes en conséquence.

Notre approche comporte un ***sélecteur de règles***, qui, après avoir appliqué la procédure de ré-écriture de règles, permet de réaliser les deux actions suivantes:

- Sélection du sous-ensemble de règles nécessaire et suffisantes pour répondre à la requête SPARQL considérée.
  - La phase de pré-traitement réécrit les clauses de Horn, en créant deux jeux de règles: l'un représente les spécialisations de concepts (RSCS), l'autre comprend les hiérarchies de propriétés (RSPS). Ces deux sous-ensembles sont utilisés durant la phase de pré-traitement pour fusionner les TBox des ontologies à aligner, uniquement en termes de relations de subsomption (e.g. **`rdfs:subClassOf`** et **`rdfs:subPropertyOf`**). Une base de connaissances séparée est générée suite à cette fusion.
- Vérification de l'existence d'éventuels cycles parmi les règles présélectionnées:
  - Le module de ré-écriture des règles exploite les définitions précédentes des ontologies source et destination (voir section 3.3.4.2.4). Dans ce contexte, une règle canonique pour l'interopérabilité est considérée être une règle SWRL dont l'antécédent est composé d'éléments de l'ontologie source, alors que son conséquent est composé uniquement d'éléments de l'ontologie de destination. Une telle règle peut être formalisée comme suit:

$$Règle\ canonique: P_1^1 \wedge P_2^1 \wedge \ldots \wedge P_i^j \rightarrow P_k^j$$

où P_j^i représente le prédicat P_i (comme présent dans l'ontologie j) et j ≠ l

L'architecture utilisée pour notre sélecteur de règles est illustrée par la Figure 55. L'algorithme de sélection de règles (voir Figure 57) ainsi que le détail des différentes actions effectuées par le module de pré-traitement de règles sont donnés dans [CI17]. L'algorithme d'exécution de requêtes est détaillé dans [CI17] et présenté par la Figure 58.

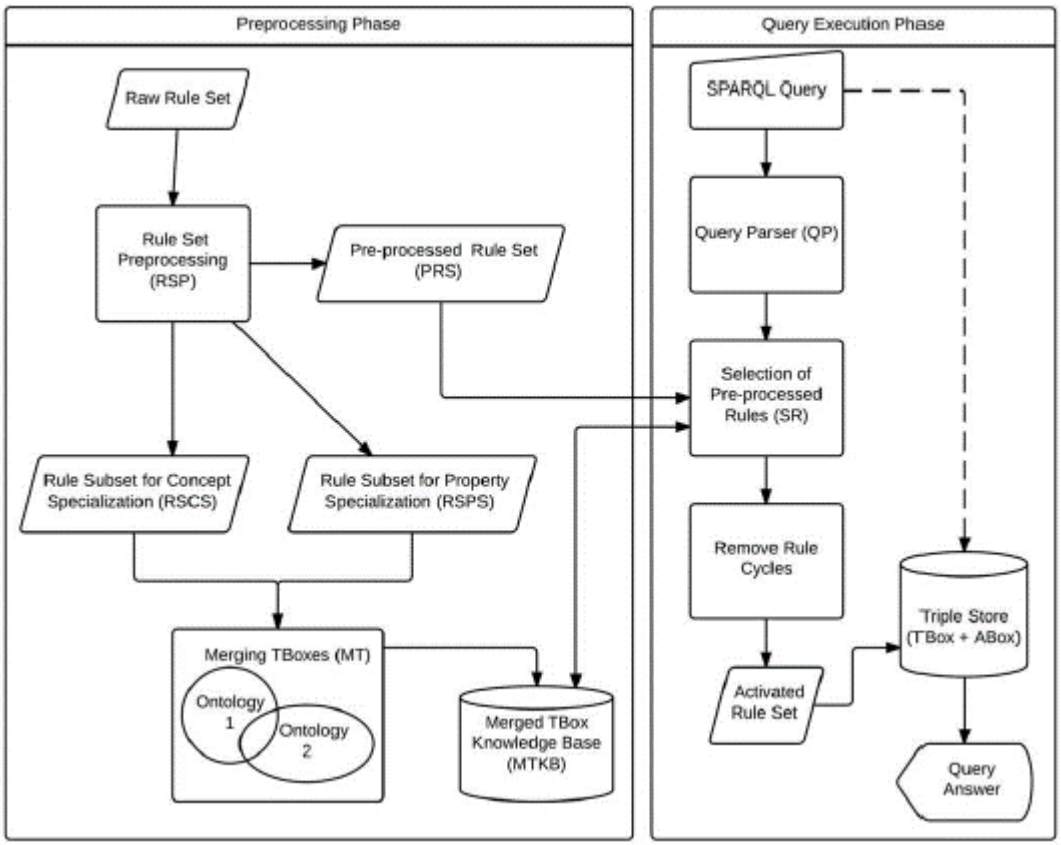


Figure 55. Architecture de notre sélecteur de règles [CI17].

La phase d'exécution de la requête proprement dite, exploite les modèles de graphe de la requête (*query graph patterns*) et les règles du jeu de règles prétraité (PRS). Elle ne considère que les règles présentes dans le jeu PRS pour l'exécution de la requête.

Le module RRC *(Remove Rule Cycles*) est chargé de supprimer les éventuels cycles restants entre les règles sélectionnées (la plupart d'entre eux ont déjà été supprimés, le module SR dans la figure ci-dessus (Figure 56) utilisant exclusivement des règles du module PRS). Pour illustrer son fonctionnement, prenons l'exemple illustré dans la Figure 56. Le module RRC identifie un cycle simple (C1=p1->p2->p1) et un cycle complexe (C2 = p1->p3->p4->p1). Pour le cycle simple, notre algorithme peut supprimer les règles R_{12} ou R_{21}. On choisit de supprimer la première règle identifiée par l'algorithme (R_{12}), mais avant de la supprimer, on matérialise dans la base de connaissances les faits que la règle permet d'inférer. Pour le cycle complexe, le module RRC peut supprimer les ensembles de règles suivants {R_{13}}, {R_{34}, R'_{34}} ou {R_{41}}. Notre algorithme supprime le premier ensemble avec la plus petite cardinalité (R_{13}), en limitant ainsi la quantité de faits à inférer (sur la base de l'hypothèse que moins un cycle contient de règles, moins il y a de triples à matérialiser).

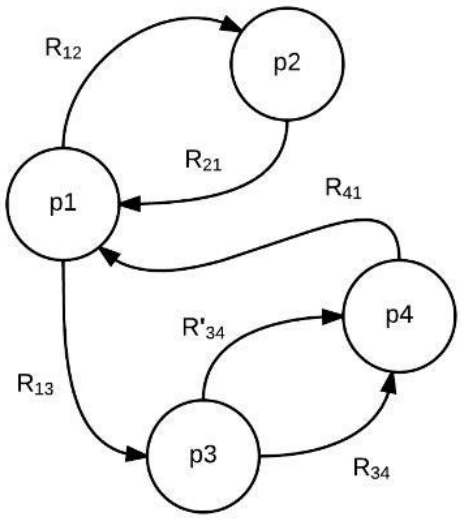


Figure 56. Exemple de cycles entre règles [CI17].

```
Input  : Raw Rule Set (R), Knowledge Base (KB)
Output : Preprocessed Rule set (PRS)
begin
    foreach vRule ∈ R do
        if vRule has not been preprocessed then
            if vRule.head contains a property with no domain/range explicitly defined
            then
                Query the KB to retrieve the domain/range information;
                if domain/range information is retrieved then
                    Rewrite the vRule with the retrieved domain/range information;
                end
            end
            vSubset = subset of R without vRule;
            if vRule can be reduced by a rule from vSubset then
                vList = create a list of lists;
                foreach vPred in vRule.body do
                    vCanonicalPredList = new List;
                    if vPred is reducible by a rule from vSubset and has a canonical form
                    then
                        vSubsetTemp = vSubset;
                        while vPred has a different canonical form do
                            vNewPred = vPred reduced using vSubsetTemp;
                            Remove from vSubsetTemp the first rule used to reduce vPred;
                            if vNewPred is a canonical form without vPred then
                                Add vNewPred to vCanonicalPredList;
                            end
                        end
                    else
                        Add vPred to vCanonicalPredList;
                    end
                end
                Create all possible rules combining elements from the different lists in
                vList;
                Add so-created rules to PRS;
            else
                Add vRule to PRS;
            end
        end
    end
end
```

Figure 57. Algorithme suivi par le module RSP pour la ré-écriture de règles [CI17].

```
Input  : Preprocessed Rule set (PRS), Knowledge Base (KB), MTKB, SPARQL Query
         (Q)
Output : Query answer (A)
begin
    U = all URIs contained in Q;
    vActivatedRules = new List;
    vRulesByHeadMap = create a map of rules from PRS where the keys are elements
    in U;
    if T_score ≥ 0 then
        foreach u ∈ U do
            vRuleList = get the rule list which u is in the rule head from
            vRulesByHeadMap;
            Calculate rule score for all rules in vRuleList

            Sort vRuleList decreasingly (according to the previous calculated score
            value);
            vActivatedRulesTemp = new List;
            Add to vActivatedRulesTemp the first T_rules rules in vRuleList with a rule score
            is equal or greater than T_score;
            if vActivatedRulesTemp is empty then
                Add all elements from vRuleList to vActivatedRules;
            else
                Add all elements from vActivatedRulesTemp to vActivatedRules;
            end
        end
    else
        Add all rules from vRulesByHeadMap to vActivatedRules;
    end
    Remove rule cycles in vActivatedRules;
    A = query KB considering the rules in vActivatedRules;
end
```

Figure 58. Algorithme pour la phase d'exécution de requêtes [CI17]

3.3.4.2.6 Évaluation de l'approche pour la fédération faiblement couplées d'ontologies OWL (C3)

Pour évaluer notre architecture de fédération d'ontologies ainsi que nos algorithmes pour la sélection de règles, nous avons choisi deux ontologies pour lesquelles nous avons défini un alignement exprimé à travers des règles SWRL. Afin de mieux illustrer les avantages de notre approche, nous avons souhaité utiliser pour ces tests un alignement comprenant de nombreuses règles. Les deux ontologies choisies présentent les caractéristiques définies dans le Tableau 29. Le lecteur pourra reconnaître les caractéristiques des ontologies COBieOWL (présentée à la section 3.3.4.2.2 COBieOWL) pour Onto1 et respectivement notre ontologie ifcOWL (présentée à la section 3.3.4.2.1) pour Onto2.

Eléments de l'ontologie	Onto1 (COBieOWL)	Onto2 (ifcOWL)
Classes	30	802
Propriétés objet	32	1292
Propriétés de type de données	125	247
Propriétés inverses	7	115
Triples dans la TBox	2503	9978
Expressivité DL	$\mathcal{ALCHIF}^{(D)}$	$\mathcal{ALUIF}^{(D)}$

Tableau 28. Caractéristiques des ontologies Onto1 (COBieOWL) et Onto2 (ifcOWL).

Notre choix s'est porté sur ces ontologies car les règles permettant de les aligner peuvent être extraites à partir du MVD COBie[25]. Ce MVD définit l'ensemble des équivalences entre les classes et les propriétés IFC 2x4 et les entités et attributs du modèle COBie. Notre ontologie ifcWoD étant basée sur la version 2x3 du standard IFC, il a d'abord fallu sélectionner uniquement les correspondances portant sur des éléments du schéma IFC 2x3. Par la suite, et sur la base de cet ensemble de correspondance, nous avons écrit un algorithme qui a généré un ensemble de 474 règles SWRL (en exploitant la sérialisation XSD du MVD). Ces règles définissent l'alignement entre le schéma IFC 2x3 et le schéma COBie. Ces règles sont exprimées en SWRL, en utilisant des éléments des ontologies ifcWoD et COBieOWL. Elles permettent de transformer automatiquement des données COBie en IFC et vice versa.

Pour l'ensemble des tests décrits ici, nous avons utilisé un magasin de triples Stardog 2.2.1, représentant le serveur, encapsulé dans une machine virtuelle avec la configuration suivante: un seul processeur Intel Xeon E5-2430 à 2,2GHz, avec 2 cœurs réservés sur les 6 disponibles, 8 Gb de mémoire RAM DD3 et une taille de 6Gb pour le tas Java de la machine virtuelle Java. Dans cette configuration, le raisonneur Stardog et le module de sélection de règles constituent notre module FC. Nous créons 4 dépôts, chacun d'entre eux contenant la base de connaissances commune aux deux ontologies considérées (e.g. TBox et ABox d'Onto1 et TBox et ABox d'Onto2). On nomme ces dépôts respectivement KB1, KB2, KB3 et KB4. Pour l'exemple décrit ici, chacun de ces dépôts contient une ABox totalisant 1 146 294 triples. A des fins d'évaluation, nous choisissons d'implémenter différents ensembles de règles dans chacun des dépôts considérés. Le Tableau 29 résumé nos différents choix. La machine client dispose d'une configuration similaire que la machine serveur: un processeur Intel Core i7-4790 à 3.6GHz avec 4 cœurs, 8Gb de mémoire RAM DDR3 à 1600MHz, et une taille de 1Gb pour la mémoire Java. Le module de sélection de règles est exécuté par la machine client.

	Nombre de règles SWRL considérées	Explications
KB1	474	L'ensemble des règles du module FLS, formant l'alignement complet entre Onto1 et Onto2.
KB2	266	L'ensemble des règles de subsomption avec l'ensemble des règles contenant des éléments d'Onto1 dans leur antécédent
KB3	178	L'ensemble des règles de KB2 moins certaines règles ayant des éléments d'Onto1 dans leur antécédent (afin de réduire le nombre d'inférences)
KB4	variable	L'ensemble des règles sélectionnées par le module de sélection de règles, dans l'ARS

Tableau 29. Bases de connaissances et configurations utilisées pour évaluer l'approche FOWLA.

[25] http://docs.buildingsmartalliance.org/MVD_COBIE/

Dans cet environnement, nous avons composé 4 requêtes SPARQL (listées dans le Tableau 30). Selon la notation utilisée dans le Tableau 30, on note C_{ij} la classe C_i dans l'ontologie i, respectivement P_l^k la propriété P_l de l'ontologie k, avec $i, j, k, l \in \mathbb{N}^*$.

Nom requête	Requête SPARQL	Nom Requête	Requête SPARQL

Q1	`SELECT ?x ?y`
	`WHERE`
	`{?x onto1:p11 ?y .}`
		Q3	`SELECT ?x ?u`
			`WHERE`
			`{?x a onto1:C11 .`
			`?y a onto2:C22 .`
			`?z a onto1:C12.`
			`?y onto2:p21 ?z .`
			`?y onto2:p22 ?x .`
			`?x onto1:p11 ?u .}`

Q2	`SELECT ?x ?y`
	`WHERE`
	`{?x a onto2:C21 .`
	`?x onto1:p11 ?y.}`
		Q4	`SELECT ?x ?y`
			`WHERE`
			`{?d onto1:p12 ?s .`
			`?c onto2:p23 ?s .`
			`?c onto2:p24 ?x .`
			`?x a onto2:C23 .`
			`?x onto2:p25 ?w .`
			`?w onto2:p26 ?y .`
			`FILTER(?y > 1.26) }`

Tableau 30. Liste des requêtes adressées aux bases de connaissances considérées (voir Tableau 29)

Chaque requête est exécutée 30 fois consécutives sur les différentes bases de connaissances considérées, à savoir KB1, KB2, KB3 et KB4. Les résultats obtenus sont affichés dans le Tableau 31 (e.g. temps d'exécution moyen et écart-type standard).

Requête	KB	Temps d'exécution moyen (en s)	Ecart-type	Nombre de règles	Nombre de résultats
Q1	KB1	-	-	474	0
	KB2	-	-	266	0
	KB3	18,04	**0,18**	178	1671
	KB4	**1,93**	1,27	**16**	**22834**
Q2	KB1	-	-	474	0
	KB2	-	-	266	0
	KB3	22,88	0,56	178	103
	KB4	**0,15**	**0,01**	**2**	**103**
Q3	KB1	-	-	474	0
	KB2	-	-	266	0
	KB3	25,19	0,42	178	2
	KB4	**0,21**	**0,02**	**4**	**2**
Q4	KB1	-	-	474	0
	KB2	0,62	0,71	266	16
	KB3	0,44	0,67	178	16
	KB4	**0,34**	**0,07**	**9**	**16**

Tableau 31. Evaluation de la performance de notre approche pour la sélection de règles.

La colonne "Nombre de règles" précise le nombre de règles sélectionnées par le module RS pour répondre à la requête considérée. Si rien n'est indiqué dans le Tableau 31 (e.g. "-") cela signifie qu'aucun résultat n'a été retourné pour la requête considérée. Ceci est souvent provoqué par un dépassement de la taille mémoire allouée à la machine virtuelle Java. Nous avons remarqué que ce phénomène se produisait environ 3 minutes après le début de l'exécution de la requête. Nous n'avons pas pris en compte

le temps de pré-traitement des différentes règles, d'une part puisqu'il n'est nécessaire que lorsque les règles sont définies pour la première fois ou modifiées, et d'autre part nous l'avons considéré négligeable. Nous avons en effet calculé le temps moyen requis pour le pré-traitement des 474 règles considérées, et nous avons obtenu une durée moyenne de 0.87s, avec un écart-type standard de 0,004 (étape de pré-traitement exécutée 30 fois sur la machine client).

Pour répondre correctement à la requête Q1, notre méthodologie a sélectionné 16 règles de l'ensemble des règles pré-traitées (PRS), définies sur la base de l'ensemble initial de 474 règles. Les résultats montrent que sans notre approche, aucun résultat n'est retourné lorsque l'on considère l'ensemble de règles PRS (en raison d'une surcharge de mémoire après environ 3 minutes d'exécution de la requête sur KB1). Lorsqu'on exécute Q1 sur KB2, toujours aucun résultat n'est retourné et ce même en réduisant les règles considérées à 266 et en supprimant les cycles entre règles. Lorsqu'exécutée sur KB3 (qui contient moins de 40% des règles de l'ensemble initial), Q1 renvoie moins de 7% des résultats attendus. Cela s'explique par le fait que plusieurs des règles pertinentes pour Q1 avaient été supprimées lors de la conception de nos bases de connaissances de test. De plus, comparé à l'exécution de Q1 sur KB4, l'exécution de Q1 sur KB3 a une durée 9 fois supérieure et récupère 10 fois moins de résultats. En effet, KB4 ne prend en compte que les règles sélectionnées (et présentes dans le module ARS). Ainsi, nous pouvons affirmer que les résultats associés à l'exécution de Q1 sur KB4 représentent le gain (en termes de temps d'exécution des requêtes et de nombre de résultats récupérés) obtenu en implémentant notre approche.

Pour l'ensemble des tests réalisés, les temps d'exécution moyens de différentes requêtes ont été considérablement réduits. Pour les tests utilisant Q2, Q3 et Q4, l'écart type pour le temps de réponse à la requête est beaucoup plus faible avec notre approche, ce qui signifie un temps de réponse plus proche de la valeur moyenne. Néanmoins, ce n'est pas le cas pour Q1 sur KB4 et KB3. Cela peut s'expliquer par le fait que Q1 renvoie moins de résultats par rapport à KB3 que par KB4, en raison de l'absence de règles dans KB3 pertinentes pour répondre correctement à Q1.

Afin d'être complets dans notre évaluation de ces contributions, nous avons souhaité comparer nos résultats à ceux obtenus avec des approches similaires. Malgré des recherches étendues, nous n'avons pas trouvé d'autres approches dans la littérature traitant de l'optimisation de requêtes SPARQL adressées sur des alignements d'ontologies exprimés par le biais de règles SWRL. Nous avons donc choisi de comparer nos résultats avec ceux obtenus avec des méthodes de ré-écriture de requêtes SPARQL, notamment les approches présentées dans (Makris 2010) et dans (Correndo et al. 2010). Pour ces tests, nous avons utilisé le même environnement technique que décrit plus haut. La ré-écriture de la requête SPARQL est effectuée par la machine client. Une fois la requête réécrite, elle est transmise à la machine serveur contenant la base de données graphe Stardog. Dans ces tests, seule la requête Q1 est considérée, puisque les autres requêtes comprennent des termes de plusieurs ontologies (et leur ré-écriture n'est pas possible avec les approches considérées). De plus, l'approche de (Correndo et al. 2010) ne permet pas la ré-écriture de Q4 à cause du mot-clé FILTER contenu dans la requête. Le Tableau 32 présente les résultats obtenus suite à 30 exécutions consécutives de Q1 sur les différentes bases de connaissances considérées.

Requête	Approche	Temps d'exécution moyen (en s)	Ecart-type	Nombre de règles	Nombre de résultats
Q1 sans raisonnement	(Makris 2010)	0,12	0,44	0	21337
	(Correndo et al. 2010)	**0,12**	**0,44**	**0**	**21337**
Q1 avec raisonnement	(Makris 2010)	33,46	0,83	0	22834
	(Correndo et al. 2010)	33,46	0,83	0	22834
	Notre approche	**1,93**	**1,27**	**16**	**22834**

Tableau 32. Comparaison des performances de notre approche avec celles de (Makris et al. 2010; Correndo et al. 2010)

Nous remarquons le fait que les deux approches considérées obtiennent des résultats identiques. Ceci n'est pas étonnant, puisque les approches considérées sont très similaires. Les résultats du Tableau 32 démontrent que lors de l'exécution de Q1 sans prendre en compte les tâches de raisonnement (par

exemple les subsomptions DL), son temps d'exécution moyen est considérablement réduit. Cependant, dans ce cas, on ne récupère pas l'ensemble des résultats (1497 résultats ne sont pas récupérés en raison de déductions manquantes).

Grâce aux tests ci-dessus, nous remarquons que Stardog implémente une mémoire cache qui permet de réduire le temps d'exécution des requêtes lorsqu'une requête est traitée plusieurs fois. La première exécution de Q1 (sans raisonnement et en utilisant les approches de (Correndo et al. 2010) ou (Makris 2010) a duré environ 2 secondes et la seconde n'a pris que 0,1 seconde . Notre approche a nécessité environ 2 secondes pour répondre à Q1 lors de la première exécution, ainsi que lors de la seconde. Cela est dû au fait que notre approche doit raisonner pour interroger les ontologies alignées. En analysant ces résultats, on peut déduire que la mémoire cache de Stardog pour les requêtes n'est pas entièrement prise en charge.

Lorsqu'on compare l'écart-type au temps d'exécution moyen pour les approches (Correndo et al. 2010) et (Makris 2010) lors de l'exécution de Q1 avec prise en charge des inférences, on remarque un écart-type bien inférieur au temps d'exécution moyen. Cela prouve que l'optimisation de la mémoire cache n'améliore pas, de manière considérable, le temps d'exécution des requêtes impliquant un raisonnement. Par rapport aux approches de (Makris 2010) et (Correndo et al. 2010), notre approche permet de récupérer les mêmes résultats avec un temps d'exécution moyen d'environ 17 fois inférieur. De plus, lorsque l'on considère la première exécution de Q1 sans raisonnement, les méthodes considérées pour la ré-écriture de Q1 ont presque le même temps d'exécution que celui obtenu avec notre approche. Finalement, notre approche réussit à récupérer l'ensemble des résultats corrects, alors que ce n'est pas le cas avec les autres approches considérées.

3.3.4.2.7 Extraction de vues IFC (C7)

Afin de répondre au problème d'extraction de sous-ensembles de données à partir d'un fichier IFC, nous avons conçu un prototype permettant d'extraire une "vue métier" à partir de tout fichier IFC. Le prototype a été programmé en Java et utilise une base de données graphe Stardog pour la persistance des données et les tâches de raisonnement associées. Pour représenter le schéma IFC, nous utilisons notre proposition ifcOWL dont les différentes métriques sont données dans le Tableau 23. La Figure 59 illustre l'architecture implémentée et ses différents modules.

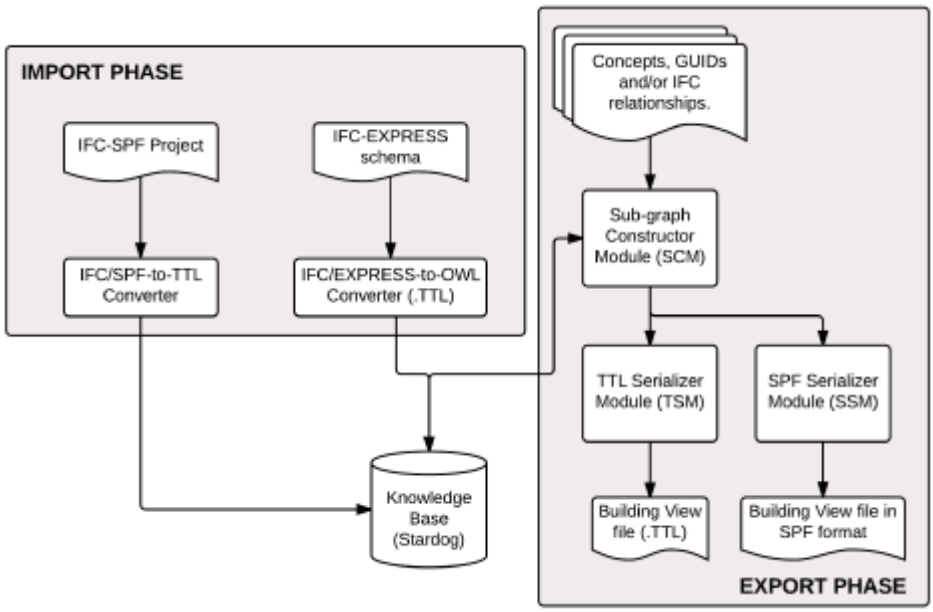


Figure 59. Architecture utilisée pour l'import et l'export des fichiers lors de la création de vues IFC.

Selon cette architecture, le processus d'extraction de vues à partir d'un fichier IFC se décompose en deux phases distinctes:

> La phase d'import concerne a) la transformation du fichier IFC-STEP considéré (représentant le projet ou le bâtiment, au format `*.spf` pour *STEP Physical File*) au format Turtle (`*.ttl`) (Allemang &

Hendler 2011), b) la sérialisation de notre ontologie ifcOWL au même format Turtle. L'ensemble schéma et données IFC est chargé dans la base de connaissances Stardog. Les processus sous-jacents de conversion et de chargement composent la phase d'import.

> La phase d'export fournit la vue sélectionnée aux formats STEP ou Turtle. La phase d'export prend en paramètre plusieurs éléments définissant la vue à extraire. Il est ainsi possible de spécifier un ensemble de concepts IFC ($\mathcal{C}$) et/ou un ensemble d'identifiants GUID ($\mathcal{G}$) et/ou un ensemble de relations IFC ($\mathcal{R}$). Ces différents paramètres composent l'ensemble Δ . Le module SCM (*Sub-Graph Constructor Module*) compose le sous-graphe de l'ensemble des instances (selon l'union des ensemble $\mathcal{C}$ et $\mathcal{G}$ fournis en entrée) avec leur propriétés. Ce sous-graphe est construit de manière récursive. Pour chaque noeud visité du graphe, le module SCM récupère la liste des propriétés spécifiées. Si la valeur de la propriété est un objet (c'est-à-dire une instance de concept), SCM récupère également cette instance et ses propriétés. Cela continue jusqu'à ce que la valeur de propriété trouvée soit un type de données (par exemple chaîne de caractères, entier). Le processus récursif présenté considère également les instances liées par le biais des relations présentes dans l'ensemble $\mathcal{R}$. Une fois le sous-graphe construit, selon le type de format choisi en sortie, le sous-graphe est envoyé soit au module TSM (*Turtle Serializer Module*) pour obtenir sa sérialisation en Turtle, soit au module SSM (*SPF Serializer Module*) pour produire sa sérialisation au format SPF.

Pour évaluer les performances de notre prototype, nous l'avons appliqué au projet d'appartement en duplex (Duplex Appartment project[26]) produit lors d'une compétition de conception architecturale, dans la ville de Weimar, en Allemagne. La Figure 60 illustre le bâtiment original, affiché avec le logiciel Solibri Model Viewer.

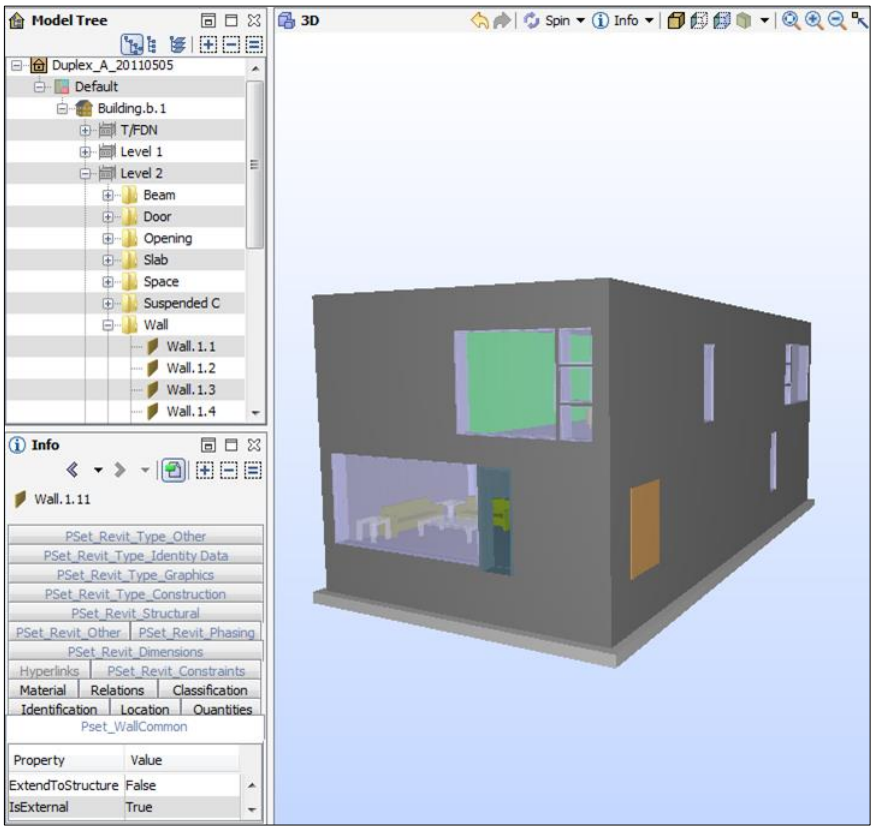


Figure 60. Visualisation avec Solibri Model Viewer du modèle bâtiment utilisé.

Sur la base de ce fichier IFC, nous avons défini le concept d' "Enveloppe du bâtiment" puis extrait la vue associée en utilisant notre prototype pour l'extraction de vues IFC. L'enveloppe d'un bâtiment représente ici l'ensemble de portes, murs et fenêtres extérieurs au bâtiment. Nous avons défini le concept d' "enveloppe du bâtiment" en utilisant des règles SWRL pour la définition des concepts sous-jacents, à

[26] Le fichier IFC correspondant à ce bâtiment est disponible en ligne: http://openifcmodel.cs.auckland.ac.nz/Model/Details/274

savoir "porte extérieure", "mur extérieur" et "fenêtre extérieure" (respectivement **:BimExternalDoor**, **:BimExternalWall** et **:BimExternalWindow** dans les équations ci-dessous). L'ensemble de ces concepts sont ensuite regroupés dans la définition d'une enveloppe extérieure d'un bâtiment (voir équations 16 à 19).

ifc:IfcDoor(?c) ^ ifc:HasProperties(?a, ?x) ^ ifc:NominalValue(?x, ?z) ^ ifc:Name(?x, ?y) ^ ifc:relPropertyDefinition(?b, ?a) ^ ifc:relObjects(?b, ?c) ^ :hasBooleanValue(?z, true) ^ :hasStringValue(?y, "IsExternal")	→	:BimExternalDoor(?c)	Equation 16
ifc:IfcWindow(?c) ^ ifc:HasProperties(?a, ?x) ^ ifc:NominalValue(?x, ?z) ^ ifc:Name(?x, ?y) ^ ifc:relPropertyDefinition(?b, ?a) ^ ifc:relObjects(?b, ?c) ^ :hasBooleanValue(?z, true) ^ :hasStringValue(?y, "IsExternal")	→	:BimExternalWindow(?c)	Equation 17
ifc:HasProperties(?a, ?x) ^ ifc:NominalValue(?x, ?z) ^ ifc:Name(?x, ?y) ^ ifc:relPropertyDefinition(?b, ?a) ^ ifc:relObjects(?b, ?c) ^ ifc:IfcWall(?c) ^ :hasBooleanValue(?z, true) ^ :hasStringValue(?y, "IsExternal")	→	:BimExternalWall(?c)	Equation 18
:BimExternalWall(?c)	→	:BimBuildingEnvelope(?c)	Equation 19
:BimExternalWindow(?c)	→	:BimBuildingEnvelope(?c)	
:BimExternalDoor(?c)	→	:BimBuildingEnvelope(?c)	

Par rapport à ces considérations, les entrées considérées par notre prototype sont donc les suivantes:

> Le fichier[27] IFC STEP Duplex_A_20110505.ifc
> L'ensemble **C** = {**BimBuildingEnvelope**} concept précédemment défini au travers des différentes règles SWRL listées ci-dessus
> L'ensemble des relations suivantes **R** = {**IfcRelDecomposes**, **IfcRelContainedInSpatialStructure**, **IfcRelVoidsElement**, **IfcRelFillsElement**, **IfcRelDefinesByProperties**}
> Aucun identifiant GUID n'est fourni - l'ensemble **G** correspond donc à l'ensemble vide

Sur la base de ces éléments, notre prototype permet de générer le fichier IFC STEP affiché dans la Figure 61 et contenant l'enveloppe extérieure du bâtiment illustré à la Figure 60.

La figure ci-contre (Figure 61) montre également ce qui semble être un mur "interne" (affiché en vert). Toutefois, ce mur possède bien la propriété IsExternal du Pset_WallCommon[28] avec une valeur "**TRUE**" dans le modèle IFC d'origine de l'appartement Duplex (voir la Figure 60). Selon la norme IFC, la valeur de la propriété **IsExternal** indique si l'élément est conçu pour être utilisé à l'extérieur (**TRUE**) ou non (**FALSE**). Si cette propriété est définie comme vraie, cela veut dire que l'élément est un élément externe, présent à l'extérieur du bâtiment. Par conséquent, cela explique pourquoi la vue de l'enveloppe du bâtiment (voir Figure 61) contient ce mur "interne". Enfin, le changement de couleurs entre la Figure 60 et la Figure 61 s'explique par le fait que les couleurs des éléments de construction ne sont pas présentes dans le modèle IFC, mais sont automatiquement définies par l'outil Solibri Model Viewer.

[27] Disponible en ligne: http://openifcmodel.cs.auckland.ac.nz/Model/Details/274

[28] http://www.buildingsmart-tech.org/ifc/IFC2x3/TC1/html/psd/IfcSharedBldgElements/Pset_WallCommon.xml

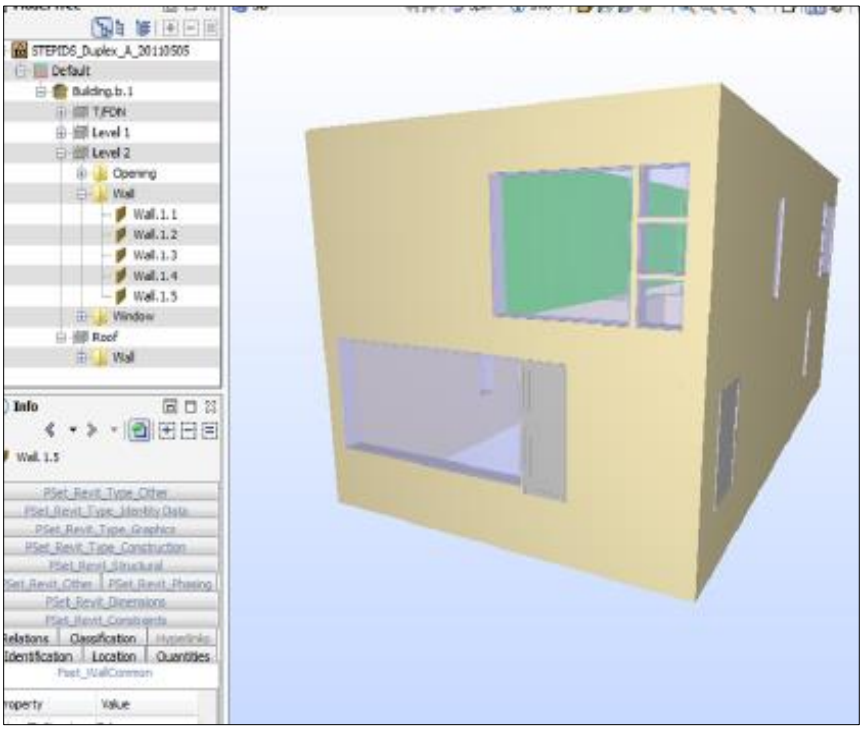


Figure 61. Affichage dans Solibri Model Viewer du fichier IFC généré avec notre approche et contenant l'enveloppe du bâtiment initial.

3.3.4.3 Rayonnement scientifique

Au total, le rayonnement scientifique associé aux recherches présentées dans ce chapitre a donné lieu à:

> 2 articles publiés dans des Revues Internationales Sélectives à Comité de Lecture [RIS1], [RIS5]
> 8 communications à des congrès internationaux à comité de sélection et actes publiés [CI8], [CI10], [CI11], [CI12], [CI13], [CI15], [CI17], [CI18]
> 1 chapitre d'ouvrage [CH7]

A côté de ces publications, plusieurs rapports de recherche internes au projet (et fournis à l'entreprise partenaire) ont aussi été rédigés. Je donne ci-dessous une liste non exhaustive des rapports produits:

> Moteurs 4D supportant l'affichage de fichiers IFC dans un environnement Web - mars 2015 - 9 pages
> Approche C-DMF et projet SIGA3D (pour notamment leur étude du standard IFC) - juillet 2014 - 16 pages
> Standard ISO 12006-3 IFD (*International Framework for Dictionnaries*) - mai 2014 - 9 pages
> IFC et Web sémantique - modélisations sémantiques pour les données IFC, pistes et approches - octobre 2013 - 12 pages
> Standard ISO 29481-3 MVD (Model View Definition) - avril 2014 - 6 pages
> Étude de l'approche COBie - août 2014 - 15 pages
> Les technologies du Web sémantique - novembre 2013 - 23 pages
> Niveaux d'expressivité en OWL - novembre 2013 - 13 pages
> Mise à jour d'ontologies - janvier 2014 - 17 pages
> IFC et ontologies - état de l'art des approches utilisant des ontologies pour la manipulation de données IFC - avril 2014 - 20 pages
> Vers une nouvelle modélisation de l'ontologie IFC - proposition d'une adaptation sémantique au niveau de la structure de l'ontologie (ifcWoD) - avril 2014 - 18 pages
> Rapport technique comparant différents magasins de triples - novembre 2013 - 17 pages
> Rapport technique sur l'algorithme de sélection de règles - février 2015 - 23 pages
> Rapport technique sur l'architecture FOWLA - mai 2014 - 11 pages

Ces rapports ont été transmis à l'entreprise partenaire et ont fait partie des transferts de compétences réalisés. Dans le cadre de la collaboration avec l'entreprise partenaire Active3D, je suis intervenue dans deux formations (dispensées à l'entreprise): formation aux technologies sémantiques (langages de description d'ontologies, langage de règles SWRL et langage de requête SPARQL) et formation à Stardog (stockage d'une base de connaissance, exécution de requêtes, possibilités de raisonnement, intégration de données, ajout de règles SWRL à une base de connaissances existante).

3.3.5 Conclusions et ouvertures

3.3.5.1 Conclusions

Aujourd'hui, les modèles de systèmes d'information d'entreprise (EIS) sont de plus en plus omniprésents et complexes. Les hypothèses à la base des recherches présentées dans ce chapitre furent de dire que 1) l'interopérabilité dans le partage d'informations entre systèmes différents est importante et 2) doit s'appuyer sur des modèles sémantiques formels (ontologies). Sur la base de ces hypothèses, notre approche FOWLA peut être appliquée à des modèles d'entreprise et/ou à des systèmes à base d'ontologie, dont l'interopérabilité est implémentée au travers de règles de Horn. Notre méthodologie de sélection des règles réduit considérablement le temps d'exécution des requêtes SPARQL adressées à ce type de système. En pratique, la mise en œuvre de l'architecture FOWLA dans des systèmes d'entreprise à base d'ontologies, permet le partage et le traitement dynamique des données, et ce directement au niveau de la base de connaissances, tout en respectant les contraintes d'exécution associées à ces environnements industriels. Les principaux avantages associés à l'architecture FOWLA ont été discutés et quantifiés. Par rapport à des approches exploitant uniquement la ré-écriture des requêtes SPARQL (sans effectuer de raisonnements), le nombre d'avantages de FOWLA est moindre, mais non sans importance. En effet, FOWLA permet d'inférer de nouveaux alignements, de composer des requêtes SPARQL utilisant des termes de plusieurs ontologies, tout en réduisant le temps d'exécution de la requête et en tenant compte des politiques d'accès aux données. Or, ceci n'est pas possible avec les approches de ré-écriture de requêtes étudiées e.g. (Correndo et al. 2010) et (Makris 2010).

Nous avons évalué et validé les performances de FOWLA dans un contexte BIM, notamment en choisissant de fédérer deux ontologies, une basée sur la norme IFC, l'autre basée sur la norme COBie. Nous avons également adapté l'ensemble des règles de correspondance entre les deux modèles sous la forme d'un alignement totalisant 474 règles SWRL. L'ontologie utilisée pour la norme IFC exploite notre processus (décrit dans la section 3.3.4.2.1) et correspond à notre ontologie ifcOWL (décrite dans la même section).

Pour pallier aux problèmes identifiés par rapport à l'approche bSI pour définir une ontologie du standard IFC, nous avons défini l'ontologie IfcWoD (voir section 3.3.4.2.3). ifcWoD exploite l'ensemble des contraintes de modélisation imposées par la structure orientée objet du schéma IFC. Par conséquent, notre contribution va au-delà d'une simple adaptation syntaxique, en proposant une adaptation sémantique du modèle IFC. IfcWoD propose une meilleure représentation des relations entre les objets d'un bâtiment et les propriétés de ces objets. En effet, la majorité des propriétés d'objets IFC sont représentées en tant qu'instances de classes de l'ontologie ifcOWL bSI, au lieu de propriétés précédemment définies dans ifcOWL. En outre, la structure sémantique du modèle IfcWoD permet de répondre également à des problèmes de performances, en optimisant le temps d'exécution des requêtes (optimisation prouvée à travers les tests présentées à la section 3.3.4.2.6).

Pour adresser le problème de l'extraction de vues IFC, nous avons développé un prototype exploitant notre ontologie du standard IFC ainsi que des règles logiques de type SWRL (pour la définition de vues et donc l'enrichissement sémantique du modèle IFC). Comme illustré à travers nos différentes publications [RIS1], notre approche à base de règles est plus souple que l'alternative la plus proche à savoir l'approche MVD de bSI (qui ne considère que le schéma, et non pas les instances). De plus, cette approche a été choisie et discutée dans le livre blanc produit par bSI (et à la rédaction duquel j'ai participé) traitant des évolutions futures de l'approche MVD. Les différentes conclusions de ce rapport ont été présentées lors du sommet international bSI qui a eu lieu à Barcelone en avril 2017, dans le cadre d'une présentation que j'ai effectuée conjointement avec Matthias Weise (AEC3).

Enfin, nous avons souhaité comparer les performances de notre approche d'enrichissement sémantique à base de règles avec des approches similaires d'autres chercheurs (approches supportant les règles SWRL, permettent le raisonnement lors de l'exécution de requêtes et utilisant des magasins de triples à grande

échelle e.g. *large-scale triple store*). Nous avons ainsi collaboré avec deux autres équipes de chercheurs internationaux et proposé deux publications comparant nos approches [RIS3][CI7]. Toutefois, les approches considérées sont bien différentes entre elles, avec des performances tout aussi hétéroclites. Il est donc difficile de faire une comparaison sincère et correcte entre ces approches, mais nous pouvons identifier des tendances générales, comme le besoin de séparer le processus de raisonnement de l'exécution de la requête proprement dite (alors que ces processus sont difficilement séparables dans le cas de Stardog et de Jena combiné avec SPIN). Cela nous a donné l'idée d'utiliser, pour l'évaluation de notre approche FOWLA, l'ensemble des modèles BIM (369), des règles (69) et des requêtes SPARQL (60) considérés pour cette étude comparative.

Concernant l'approche FOWLA, les travaux futurs prévus concernent l'extension de son évaluation en utilisant d'autres moteurs d'inférence tels F-logic (Kifer 2005), ObjectLogic (Angele 2014), RIF ((W3C) 2013), SPIN (Knublauch 2011) ou encore N3Logic (Berners-lee et al. 2008). Nous souhaitons aussi adapter notre prototype d'extraction de vues IFC afin qu'il puisse utiliser l'architecture FOWLA pour permettre la fédération de sources de données hétérogènes lors de la construction de vues métier. Selon la taille du fichier IFC considéré en entrée et la complexité de la vue à extraire, ceci peut être une tâche fastidieuse en termes de temps d'exécution et ressources impliquées. Au niveau du sélecteur de règles, il sera important d'étudier les performances d'une approche basée sur le formalisme RIF défini par le W3C, au lieu de règles SWRL. Un dernier point, mais pas des moindres, concerne l'implémentation d'un mécanisme pour vérifier si l'introduction de nouvelles règles ne génère pas des incohérences ou des ambiguïtés. Pour l'instant notre approche exploite le mécanisme de Stardog pour la détection de telles incohérences.

Par rapport à ifcWoD, les travaux futurs concernent l'analyse du compromis à faire entre redondance des données et performance des requêtes SPARQL. En effet, la manière dont sont définis les domaines des différentes propriétés contenues dans les *Property Set Definitions* (PSD) impacte le temps d'exécution des requêtes SPARQL les utilisant. Il faudrait réaliser des tests supplémentaires afin de déterminer avec précision s'il vaut mieux modéliser ces domaines en tant que classes IFC au lieu de classes `IfcPropertySet`. De plus, il existe plusieurs propriétés (e.g. `reference`) qui se retrouvent à l'intérieur de plusieurs PSDs. Il faudrait étudier davantage le recueil et la hiérarchisation de ces propriétés. Ces travaux pourront être mis en lien avec les travaux de Maxime Lefrançois visant l'utilisation de primitives SPARQL Construct pour les PSDs.

3.3.5.2 Ouvertures

Les travaux présentés dans ce chapitre ont donné naissance à de nombreuses ouvertures et collaborations.

Au niveau international, les articles comparant notre approche avec celles d'autres chercheurs ont été rédigés de manière collaborative (via le système en ligne overleaf). Cette collaboration a fait naître des réflexions ultérieures sur:

> La ***structure de l'ontologie ifcOWL de bSI*** - ce qui a permis de rédiger un article co-signé par moi-même et Dr. Ing. Arch. Pieter Pauwels [CH4] où nous proposons plusieurs adaptations du schéma de l'ontologie ifcOWL, dont l'adaptation d'IfcRelationship en tant que propriété (inspirée de notre approche ifcWoD)
> La ***modélisation des processus métiers avec la norme*** IDM - ce qui a résulté en un article rédigé avec Dr. Udo Kannengiesser où nous proposons une alternative à la représentation standard (basée sur BPMN) basée sur le langage S-BPM (Subject-oriented Business Process Modelling) [CH5]

Dernièrement, j'ai été invitée à faire partie du jury de thèse de Monsieur Chi Zhang, encadré par Prof. Jakob Beetz.

Au niveau national, les différentes présentations que j'ai fait de nos travaux lors de sommets internationaux bSI m'ont permis d'établir des contacts et des projets avec les acteurs suivants:

> CSTB (Centre Scientifique et Technique du Bâtiment) - dans le cadre d'une prestation tarifée, j'ai été en charge d'une formation à l'écriture de règles SWRL et leur intégration dans un environnement Stardog. Cette formation a été dispensée dans les locaux du CSTB à Sophia Antipolis, sur une durée de 3j, en août 2017. J'ai par la suite participé à la rédaction de deux articles scientifiques, récents, avec des experts du CSTB - [CI1], [CH1]. Nous avons comme projet de déposer une réponse à l'appel

à projets ANR Blanc 2018. Nous souhaitons démarrer des travaux de thèse autour de la vérification de maquettes numériques par le biais de règles logiques, et ce à partir du début 2019.

- eGIS - Depuis le mois de juin 2018, j'encadre, d'un point de vue scientifique, les travaux de thèse mastère professionnel de Madame BARRUE-BASSIN (experte sénior SIG chez eGIS). En collaboration avec Christophe CASTAING, son encadrant en entreprise et responsable du programme Ingénierie Numérique chez eGIS, nous avons étudié une approche pour implémenter une interopérabilité entre données BIM et données SIG. La soutenance de cette thèse est prévue pour la fin de cette année 2018.
- Nobatek/INEF4 - Avec le Prof. Nicolle, nous avons été sollicités par Nobatek afin de co-encadrer un contrat postdoctoral visant la conception d'un système pour la fédération de données urbaines.
- PTNB (Plan Transition Numérique Bâtiment) - Sous l'impulsion du CSTB, j'ai participé à la rédaction de la feuille de route "Normalisation" du PTNB, identifiant les approches de pré-normalisation et de normalisation prenant place au niveau français, mais également au niveau européen et national, et ce dans le domaine des données liées et au niveau des approches MVD. J'ai ainsi rédigé deux chapitres (un sur la qualité des vocabulaires de données liées existant, et un sur les conteneurs de données), qui ont été résumés dans le document publié par le Ministère du Logement en avril 2017[29].
- Cerqual / Qualitel - Suite à ma présentation lors du salon BIM World, à Paris, en mars 2017, j'ai débuté une collaboration avec l'organisme Qualitel (responsable en France des normes NF et NF Habitat). L'idée est d'étudier les possibilités existantes pour la vérification semi-automatique de maquettes numériques par rapport aux contraintes spécifiées dans les différents DTU sous-jacents à l'attribution de normes NF.

Au niveau des différents organismes de standardisation où j'interviens, j'ai été sollicitée et j'ai intégré les différents groupes de discussion ci-dessous:

- ISO
  - Membre du groupe de travail joint entre l'ISO/TC 59/SC 13 et l'ISO/TC 211 sur l'interopérabilité BIM et SIG (ISO TR 23262 "GIS (Geospatial) / BIM interoperability")
  - Sécurité des échanges informatiques dans un contexte de chantier - ISO/DIS 19650 "Organization of information about construction works — Information management using Building Information Modelling (BIM)" (Part 5: Specification for security-minded building information modelling, digital built environments and smart asset management)
    - Membre du comité éditorial pour la norme ISO 21597 "Information container for data drop" (échange et intégration de données à travers les conteneurs sémantiques)
- buildingSmart International (bSI)
  - J'ai participé à la rédaction du manuel bSI pour la conception d'ontologies - j'ai été notamment en charge de rédiger la partie présentant les langages RDF Schema et OWL (1 et 2, avec les différents sous-profils), de même que la partie expliquant le raisonnement avec les approches sémantiques. Finalement j'ai fourni une grande partie du texte composant la section présentant les approches Linked Data (ou données liées).
  - J'ai participé à la rédaction du livre blanc bSI discutant de l'avenir de l'approche MVD. Notre approche consistant à interpréter le MVD sous la forme de règles SWRL (telle qu'implémentée dans l'évaluation de l'architecture FOWLA) a été identifiée comme pertienete et citée parmi les trois possibilités d'évolution mentionnées. Chacune des possibilités exploite une approche à base de règles logiques, confirmant ainsi notre intuition d'utiliser ce genre d'outil pour faciliter la gestion des connaissances experts.
- AFNOR
  - Membre de la commission AFNOR PPBIM "Maquettes numériques dans la construction", plus particulièrement j'interviens dans les groupes d'études GE1 "Feuille de route", GE3 "Information Delivery Manual" et GE4 "Dictionnaire de données"
  - Projet BoostConstruction - intégration de données constructeur, de données produit, de catalogues produit, de classifications, afin de créer un CDE (C*ommon Data Environment*)

[29] L'étude est disponible en ligne : http://www.batiment-numerique.fr/uploads/DOC/PTNB%20-%20FdR%20Normalisation%202017.pdf

3.4 Projets de recherche en cours et perspectives

En cette fin 2018, nous nous trouvons en plein milieu de la quatrième révolution industrielle. Les initiatives de "transformations numériques" que ce soit dans les entreprises publiques ou privés, ou encore au niveau des territoires, deviennent monnaie courante de nos jours – dès la fin 2017, plus de 70% des plus grandes entreprises mondiales (en termes de revenus[30]) avaient des équipes en charge d'assurer la transformation numérique / l'innovation. Ce nombre devrait être multiplié par 2 voire 3 d'ici 2020. Nous vivons un réel changement de paradigme au niveau des entreprises IT : nous passons de systèmes à des services, d'une agilité informatique à une agilité métier, et enfin au lieu de parler d' "information" nous parlons d' "innovation".

Souvent décrite comme le résultat de l'intégration et des effets combinés de plusieurs "technologies exponentielles», telles que l'intelligence artificielle (IA), les biotechnologies et les nanomatériaux, la 4e Révolution Industrielle étend le paradigme de la révolution industrielle à un avenir où beaucoup d'éléments de ce que nous considérons "industrie" (e.g. usines fixes et centralisées, main-d'œuvre massive au sein des grandes entreprises) n'existeront plus. Un exemple de la vision émergente de cette 4e Révolution est le développement d'organismes synthétiques (créés à partir d'ADN conçu par ordinateurs et bio-imprimé) fabriqués à l'aide de chaînes de montage robotisées, où les nanomatériaux améliorent énormément l'efficacité de la production. La 4e révolution industrielle a dévoilé l'apprentissage automatique, les robots, les machines intelligentes, l'impression 3D, l'Internet des objets (IoT pour *Internet of Things*) et encore les systèmes cyber-physiques (CPS pour *Cyber-Physical Systems*). Les "prouesses technologiques" les plus connues sont l'augmentation exponentielle de la puissance de calcul des ordinateurs et la diminution des coûts de stockage, qui obéissent à la loi de Moore. Le doublement de la puissance de calcul des processeurs tous les 18 à 24 mois a permis à de nouveaux supercalculateurs, e.g. Milky Way 2 (Peters 2017), d'atteindre des vitesses de calcul de 300 quadrillions de FLOPS (*FLoating Operations Per Second* ou opérations flottantes par seconde). En seulement deux décennies, la puissance de calcul a été augmentée avec un facteur supérieur à 300 000 ! Toutefois, lorsque ces technologies numériques sont combinées à d'autres technologies en expansion rapide, telles que la biotechnologie, la nanotechnologie et l'IA, le rythme du changement est d'autant plus élevé. Cette convergence des technologies a été qualifiée de «singularité», et, selon certains auteurs, devrait procurer à l'humanité des bienfaits indicibles (Kurzweil 2005).

Le Forum Mondial Economique (*World Economic Forum* ou WEF) a défini un ensemble de points de basculement au cours desquels les technologies de la 4e Révolution Industrielle se généraliseront suffisamment pour créer un changement sociétal massif. Ces points critiques incluent la prolifération des technologies de la 4e Révolution à des niveaux où elles ont un impact significatif sur nos vies et nécessitent des changements dans l'emploi et l'éducation. Une enquête (Forum 2015) menée auprès de 800 experts et cadres high-tech a déterminé une série de dates auxquelles ces points de basculement seraient atteints. Les exemples incluent la pose d'implants pour téléphones portables en 2025, 80% des personnes ayant une présence numérique d'ici 2023, 10% des lunettes de lecture connectées à Internet d'ici 2023, 10% des personnes portant des vêtements connectés à Internet d'ici 2022, 90% de la population mondiale ayant l'accès à Internet d'ici 2024, 90% de la population utilisant des smartphones d'ici 2023, 1 000 milliards de capteurs connectés à Internet d'ici 2022, plus de 50% du trafic Internet occupé par les applications domotiques et les appareils ménagers d'ici 2024 ou les voitures sans chauffeur qui vont représenter 10% de toutes les voitures aux États-Unis d'ici 2026. De nombreuses autres prévisions suggèrent une intégration poussée de l'IA au travers de pharmaciens robotisés, de la prolifération des bitcoins dans l'économie, les véhicules imprimés en 3D d'ici 2022 ou encore les transplantations d'organes (e.g. le foie) imprimés en 3D d'ici 2024.

Ces développements façonnent une nouvelle économie de données. En tant que nouvelle matière première essentielle pour la série de développements macro-économiques discutés ci-dessus (voir Figure 62), ces ressources de données, bien qu'abondantes et omniprésentes, constituent les entrées de cette révolution. Elles viennent s'ajouter aux ressources communément utilisées pour la production, classées selon (Arthur O'Sullivan 2016) en trois catégories et toujours en train de se raréfier face à des besoins humains en continue augmentation.

1. Ressources naturelles: terre, minéraux, pétrole, gaz naturel, eau, etc.

[30] http://fortune.com/global500/

2. Ressources en capital: machines, équipements durables, outils, infrastructures, bâtiments, etc.
3. Ressources humaines: travail physique, effort intellectuel, connaissances, compétences, expériences, leadership, esprit d'entreprise, etc.

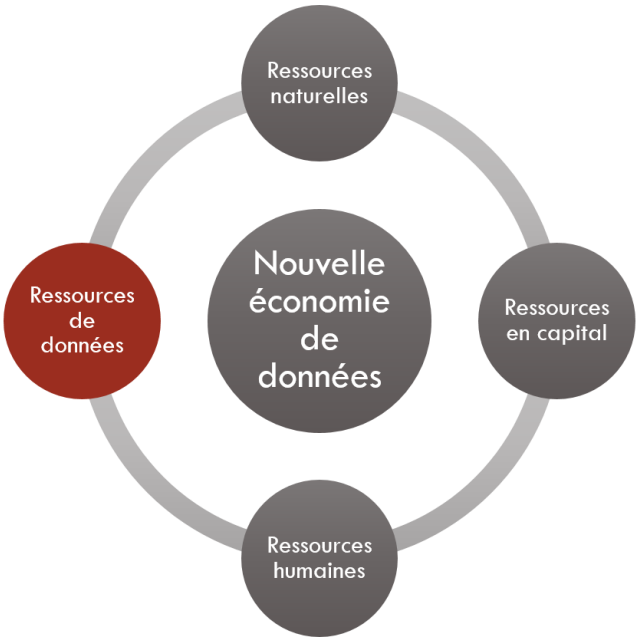


Figure 62. Les ressources de données entrent dans la nouvelle économie de données

Le démarrage d'une voiture autonome peut générer environ 100 giga-octets de données en une seconde. Certaines personnes considèrent les données «Big Data» en tant que «données massives». Cependant, la définition la plus utilisée demeure celle des "Trois Vs" (Yaqoob et al. 2016). Elle a été proposée par Doug Laney, analyste éminent chez Gartner: les données "Big Data"" sont généralement définies comme "des informations à grand volume, à grande vitesse et / ou de grande variété nécessitant de nouvelles formes de traitement pour permettre une optimisation des processus de prise de décision et de découverte de connaissances[31]".

Sans accès aux données et sans une utilisation optimale, l'apprentissage automatique et l'IA ne mèneront pas à de nouveaux développements. Les données s'apparentent à de la matière première. Elles permettent aux gens de voir au-delà du monde numérique et ouvrent la voie à des communications plus rapides (conception centrée sur le client), à des mécanismes de coordination améliorés (logistique intelligente et usines intelligentes) et à des collaborations innovantes (par exemple, nouveaux modèles commerciaux et de coopération à échelles régionales voire continentales).

Les entreprises d'aujourd'hui ont augmenté leur capacité à analyser et à utiliser les données massives. Par exemple, les entreprises Internet gèrent des données qui leur confèrent un énorme pouvoir (e.g. Alphabet, Amazon, Apple, Facebook et Microsoft) (Economist 2017b), tandis que les géants industriels et les détaillants collectent des données qui les rendent plus compétitifs (e.g. Siemens, GE, McDonald et Tesla) (Barton & Court 2012). Les données seraient un moteur de croissance et de transformation, la véritable main invisible derrière la quatrième révolution industrielle. Les données sont en train de devenir un moyen pour les gouvernements, les secteurs privés et les organisations d'améliorer la précision et la fiabilité (IBM 2014), tout en offrant des perspectives cachées susceptibles de façonner et d'influencer les environnements de fabrication, sociaux et de service (Economist 2017a). Par conséquent, l'exploitation de la puissance du Big Data déterminera les chances de succès des entreprises à l'avenir (Oguro 2016). À l'instar des sources d'information précédentes, on pense que les données favorisent les néo-infrastructures, les néo-entreprises, les néo-monopoles, la néo-politique et, surtout, les nouveaux modèles / systèmes économiques. Cependant, contrairement aux sources d'informations antérieures, les données sont extraites, analysées, affinées, évaluées, achetées et vendues de diverses manières. A l'ère de cette

[31] Selon Gartner « *Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.* » (Disponible en ligne : https://www.gartner.com/it-glossary/big-data)

4e Révolution Industrielle, les droits de propriété et l'utilisation des données sont susceptibles de générer des conflits.

Bien que la 4e Révolution Industrielle soit censée générer des avantages non-exclusifs, elle pourrait également poser problème à notre société et entraver le développement (social / socio-économique) en remplaçant le capital humain par des machines. La main-d'œuvre doit donc avoir un meilleur accès à l'enseignement supérieur et à une éducation de qualité. Le travail peu qualifié deviendra plus rare à l'avenir, car les tâches prévisibles seront remplacées par des machines. Les travailleurs hautement qualifiés comme les moins qualifiés doivent être formés différemment. Le développement rapide des diverses technologies a conduit à l'automatisation partielle ou totale de nombreux métiers. Bien que nombre d'entre nous s'inquiètent de la possibilité d'une situation dans laquelle le travail humain est remplacé par l'automatisation et les technologies numériques, il est également important de reconnaître que l'automatisation des tâches peut être un changement positif, du moins dans les cas suivants (Phillips & Phillips 2015):

1. Les emplois se caractérisent par la monotonie et l'ennui: ces types d'emplois sont généralement routiniers et exigent plus de concentration que de pensée critique (par exemple, les postes de la chaîne de montage). Les humains ont tendance à se sentir insatisfaits lorsqu'ils occupent des emplois monotones et ennuyeux, ce qui pourrait entraîner un absentéisme, un roulement élevé du personnel, des accidents voire même une détérioration de la santé.
2. Les emplois sont caractérisés par une dangerosité défavorable: de tels emplois sont courants dans des secteurs comme la fabrication, l'exploitation minière, l'énergie nucléaire ainsi que dans d'autres industries lourdes. L'introduction de l'automatisation pourrait prévenir les blessures et les décès inutiles tout en augmentant la productivité.
3. Les métiers impliquent des transactions simples: les guichets automatiques bancaires sont des exemples dans cette catégorie, principalement parce qu'ils fonctionnent 24 heures sur 24, 7 jours sur 7. Des technologies avancées similaires pourraient rendre les transactions à étapes plus efficaces que lorsqu'elles sont gérées par des opérateurs humains conventionnels.
4. Emplois indésirables pour les humains: Il est devenu de plus en plus difficile de recruter des membres d'équipage qualifiés prêts à être en mission pendant des mois. Lors du transport de denrées non périssables, les cargos autonomes sont plus pratiques car ils permettent d'économiser sur les frais d'hébergement de l'équipage et d'éliminer la majeure partie des services publics associés (chauffage, plomberie, etc.).

Compte tenu des éléments ci-dessus, il devient évident que l'investissement dans le capital humain est nécessaire à des fins multiples, comme la finalisation de la prise de décision, la résolution de problèmes et la surveillance des processus. Même dans un environnement de travail entièrement automatisé, l'homme reste indispensable. Lorsque de nouvelles technologies sont introduites pour la première fois, les humains sont nécessaires pour finaliser et coordonner les tâches de mise en œuvre. Lorsque des systèmes sont mis en service, les utilisateurs doivent effectuer des tâches de maintenance non triviales. Les êtres humains ont également la capacité de mettre à niveau leurs compétences en prenant le relais lorsque l'automatisation échoue. Cette analyse peut nous aider à conclure que le capital humain n'est pas dépassé à l'ère de la 4e RI, mais nécessite une formation renforcée.

En effet, lorsque l'on considère les trois premières révolutions industrielles, les machines ont surpassé les hommes en termes de tâches mécaniques. Cela a conduit à un changement dans les tâches associées au travail humain, passant de des tâches mécaniques à des tâches cognitives dans le secteur des services. Les technologies numériques de la 4e RI (voir Figure 63), avec l'avènement de l'IA, sont sur le point de surpasser les humains dans les tâches cognitives (Frey & Osborne 2013). La tendance actuelle montre que les algorithmes basés sur l'IA pour le Big Data sont en train de remplacer, sur le lieu de travail, un grand nombre de tâches cognitives inhabituelles. L'informatisation des emplois laissera une grande partie du travail humain sans emploi. Erik Brynjofsson et Andrew McAfee ont montré, dans (Brynjolfsson & McAfee 2012), qu'il existait une forte relation négative entre les salaires et le niveau d'instruction et la probabilité d'informatisation. Les futurs professionnels devront se tourner vers de nouveaux domaines et types de travail.

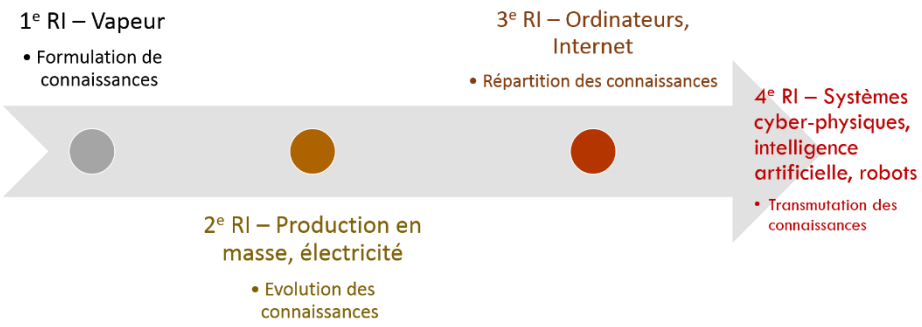


Figure 63. La quatrième révolution industrielle ou l'effacement des frontières entre les humains et les machines.

Malgré l'abondance de données et de ressources humaines qui laisse présager un bel avenir pour le nouveau monde amené par cette 4e RI, il existe un fossé entre la vérité (comment les ressources de données sont exploitées) et les faits (comment les ressources humaines parviennent à s'adapter). En effet, l'IA se base sur des processus informatisés qui imitent le comportement et le raisonnement humain dans la résolution de problèmes. Bien que les débuts de l'IA aient été pleins d'échecs, ce domaine de recherche a refait surface grâce à de nombreux chercheurs qui se sont consacrés à l'intégration de théories mathématiques rigoureuses (approches d'inférence statistiques, représentation topologique, théorie des matrices relationnelles) (Marwala & Hurwitz 2017). Du point de vue du public, l'IA semble suffisamment puissante pour déclencher cette 4e RI et résoudre automatiquement un large éventail de problèmes. En dépit du fait qu'il existe encore un énorme fossé entre la résolution de problèmes par les ordinateurs et la résolution de problèmes par les humains, les scientifiques s'efforcent de le combler (Xing & Marwala 2018). Selon Hobbs (Hobbs 1985), la capacité de ***conceptualiser le monde à différents niveaux et de bénéficier d'une mobilité totale entre ces niveaux est une caractéristique exclusive de la résolution humaine de problèmes***. En effet, lorsque nous regardons le monde qui nous entoure nous n'en retirons que les choses qui servent nos intérêts du moment.

Par exemple lorsque nous planifions un voyage, il suffit de d'approximer la route sous la forme d'une courbe unidimensionnelle. Par contre, lorsque nous traversons une route, nous la considérons comme une surface et ce n'est que lorsque nous devons y creuser des trous, qu'elle devient un volume pour nous. Lorsque nous conduisons sur une route, nous alternons ces granularités, parfois conscients de notre progression le long de la courbe unidimensionnelle, effectuant d'autres fois des ajustements dans notre position sur la surface de la route, et enfin à d'autres moments, en ralentissant devant d'éventuelles bosses ou nids de poule (dans le volume de la route). Ce trait est la marque de l'intelligence humaine.

Cependant, lorsque nous portons notre attention sur nos homologues, les machines intelligentes, la situation change radicalement. Bien que les ordinateurs disposent de presque toutes les données relatives à une usine (conditions de machine, profils de travail, disponibilité des outils, environnement de construction et liste d'inventaire), il leur est toujours difficile de créer différents modèles de représentation sur la base de ces données, sans parler de la possibilité de basculer entre des modèles avec différents niveaux de granularité.

À cet égard, les humains ont déjà entrepris de résoudre ce problème en utilisant leur propre intelligence. Bien que l'implémentation physique de machines intelligentes soit probablement très différente de celle des cerveaux humains, une compréhension approfondie des règles de travail fondamentales du cerveau humain demeure une exigence souvent indispensable pour concevoir des machines utilisant l'IA. Sur la base de divers efforts, l'***informatique granulaire*** apparaît comme un candidat approprié pour satisfaire aux exigences de représentation et de modélisation de l'intelligence humaine. L'idée fondamentale de l'informatique granulaire est inspirée par la capacité de l'homme à traiter l'information de manière granulaire, à plusieurs niveaux. Les fondements de l'informatique granulaire peuvent être retracés à la théorie des parties et des touts ou la méréologie (le mot est dérivé du mot grec "*mereos*" signifiant partie ou portion ou segment).

Théorie remontant au 6e et 5e siècles avant JC (Varzi 1996), la ***méréologie*** a été discutée et étudiée dans le travaux de Lesniewski sur le concept d'extension et la théorie formelle des noms (ou l'Ontologie pour

le calcul des noms en remplacement du calcul des prédicats) (Lesniewski 1992) ou encore ceux de Leonard et Goodman et leur approche de "calcul des individus" (basée sur la logique standard des prédicats du premier ordre) (Goodman & Leonard 1940). Sa formalisation à travers des axiomes exprimés en algèbre booléenne, exposée dans l'ouvrage "*Parts, a Study in Ontology*" de Peter Simons (Simons 1987) a donné naissance à ce que nous appelons aujourd'hui la méréologie extensionnelle classique (*Classical Extensional Mereology* ou CEM). L'approche CEM se base sur trois axiomes[32] et nécessite la définition d'une relation primitive sur laquelle peut être défini l'ensemble du système méréologique:

- Axiome 1 - ***Composition sans limites*** - Lorsqu'il existe des parties, il existe une somme méréologique de ces parties
- Axiome 2 - ***Unicité de composition*** ou PPP (*Principes des Parties Propres*) - si deux objets ont les mêmes parties propres alors ils sont identiques

$$(\forall x, y)[\big((\exists z)[(z \ll x)] \wedge (\forall z)[(z \ll x) \supset (z \ll y)]\big) \supset (x < y)]$$

- Axiome 3 - ***Transitivité*** - Si x fait partie d'une partie de y, alors x fait partie de y

$$(\forall x, y)\big[\big((x \ll y) \wedge (y \ll z)\big) \supset (x \ll z)\big]$$

Les relations primitives peuvent être choisies parmi:

- La ***définition de parties propres*** - il s'agit de relations d'ordre qui sont réflexives, transitives et antisymétriques - x est une partie propre de y est exprimé par x << y, x est une partie de y si et seulement si x est une partie propre de y ou x est identique à y

$$(\forall x, y)\ \big[(x < y) \equiv \big((x \ll y) \vee (x = y)\big)\big]$$

- Le ***chevauchement*** - Deux objets se chevauchent si et seulement s'ils ont une partie en commun (relation réflexive, symétrique et non-transitive). Dans l'équation ci-dessous, "x et y se chevauchent" est exprimé par "x o y".

$$(\forall x, y)\ [(x \circ y) \equiv (\exists z)[(z < x) \wedge (z < y)]]$$

- La ***disjonction*** - Deux objets sont disjoints si et seulement s'ils ne se chevauchent pas, c'est-à-dire s'ils n'ont pas de partie en commun. Dans l'équation ci-dessous, "x et y sont disjoints" est exprimé par "x | y".

$$(\forall x, y)\ [(x \mid y) \equiv\sim (x \circ y)]$$

- La somme binaire, le produit binaire, la limite supérieure (moindre ou moindre générale), la somme générale, le produit général, la différence, etc.

Partant de ces considérations, l'informatique granulaire exploite la théorie de la granularité (Hobbs 1985). Cette théorie suppose que le monde qui nous entoure, quel que soit le niveau de complexité considéré, du moment qu'il peut être géré par un ordinateur, peut être représenté à travers une théorie globale pouvant s'apparenter à une théorie de la logique du premier ordre $\mathcal{T}_0$. La granularité consiste à extraire de cette théorie globale $\mathcal{T}_0$, des théories locales, plus faciles à interpréter et calculer. Si l'on considère $\mathcal{P}_0$ comme l'ensemble des prédicats disponibles dans $\mathcal{T}_0$, S_0 étant le domaine d'interprétation, alors si on considère un sous-ensemble $\mathcal{R}$ de $\mathcal{P}_0$ comme étant l'ensemble des prédicats pertinents par rapport à une situation donnée, alors il est possible de définir la ***relation d'indiscernabilité*** sur S_0, $\sim$, à travers l'équation suivante (Hobbs 1985):

$$(\forall x, y)\ [(x\ \sim\ y) \equiv (\forall p \in R)(p(x) \equiv p(y))]$$

Cette relation permet de spécifier que x et y sont considérés ***indiscernables*** s'il n'y a pas de prédicat pertinent les différenciant. Déterminer les prédicats pertinents par rapport à une situation est un problème de recherche en soi. Si l'on reprend l'exemple du voyage ci-dessus, dans un tel contexte, le

[32] D'après http://encyclo-philo.fr/mereologie/

seul prédicat pertinent serait la distance du point de départ ou la distance par rapport à l'arrivée. Deux points dans l'asphalte, identifiés par leurs coordonnées par rapport au volume d'asphalte et non le long de la route, seront indiscernables. L'application de l'axiome d'indiscernabilité définit ci-dessus nous permet donc de simplifier la représentation de la route d'un volume à une courbe unidimensionnelle.

La granularité est une nouvelle approche de structuration des représentations de la connaissance, approche qui est orthogonale aux méthodes existantes de modélisation. Elle permet un découpage du domaine de connaissance ciblé selon des méthodes structurées et cohérentes, impliquant une définition des types de granularité considérés ainsi que des critères formant une perspective granulaire. Ceci peut être implémenté à différents niveaux de détail, de manière à éviter soit des enrichissements trop détaillés, soit des généralisations trop grossières. Une telle approche permet de raisonner sur l'ensemble du domaine de connaissance ainsi modélisé, ainsi que à travers des différents niveaux de granularité considérés.

La description ci-dessus est volontairement simple - l'idée n'est pas de faire un état de l'art de l'informatique granulaire, des approches et tendances dans le domaine, qui ont été très nombreuses durant les deux dernières années (1330 publications indexées sur Google Scholar en 2018, 1500 en 2017, contre seulement 4680 publications entre 2014 et 2017). Pour l'instant, il est suffisant de dire l'informatique granulaire aide les machines à mieux imiter la capacité de résolution de problèmes de l'homme.

Cette affirmation me permet avant tout d'introduire la question recherche que je poursuis dans mes travaux actuels et que j'entends poursuivre à travers mes travaux prospectifs. Cette question est la suivante: comment et à quel niveau l'usage de granularités permet d'améliorer la représentation et la gestion de connaissances ? Cette interrogation est directement issue à partir d'une analyse des limites atteintes dans mes recherches passées, notamment l'incomplétude inhérente à la connaissance et le besoin de gérer des connaissances imprécises (approché au travers du co-encadrement de Madame HOPPE), ou encore le besoin sélectionner juste les règles pertinentes par rapport à un contexte (approché au travers du co-encadrement de Monsieur MENDES de FARIAS). De plus, ces recherches passées me permettent aujourd'hui d'identifier des éléments de réponse possibles: une définition formelle du concept de granularité (et des niveaux et perspectives associés) permettrait de l'intégrer aux approches existantes et reposant sur des ontologies.

Par contre, ces quelques éléments de réponse ne permettent pas pour autant de résoudre les verrous restants. Ceux-ci sont discutés dans les paragraphes suivants. En effet, le concept de granularité, même s'il semble intuitif et facile à implémenter, semble avoir des définitions différentes selon le domaine de connaissance considéré. Alors que des modèles formels de la granularité existent depuis quelques années e.g. (Mani 1998), (Keet 2008), à la manière des ontologies, les niveaux et perspectives associés doivent être spécifiés de manière explicite et formelle en intégrant les caractéristiques du domaine de connaissance concerné (Livi & Sadeghian 2016). De plus, lorsqu'il s'agit d'intégrer des granularités dans les applications manipulant des connaissances métier et s'adaptant à un contexte (utilisateur ou d'utilisation), il convient d'investiguer le lien entre la granularité des connaissances et la granularité du contexte.

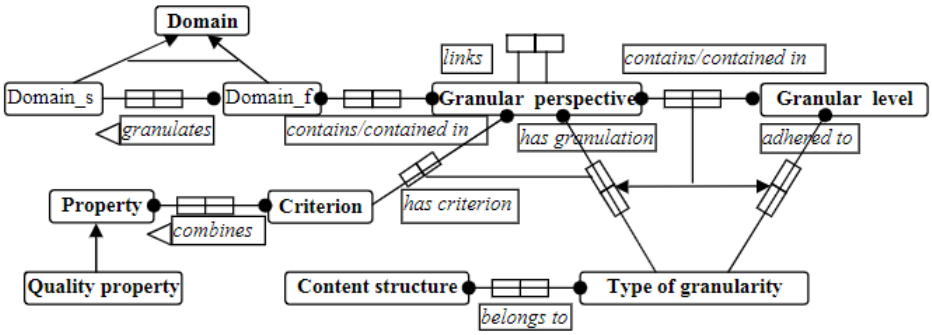


Figure 64. Perspectives et niveaux dans les théories de granularité d'après (Jakubczyc & Owoc 2011)

La granularité des connaissances apparaît en effet comme un point crucial pour l'ingénierie des connaissances (Mach & Owoc 2010). Les étapes classiques définies pour le processus général de gestion des connaissances sont les suivantes: acquisition, stockage, diffusion et intégration des connaissances dans une application. La granularité des connaissances impacte la première étape (nous pouvons recueillir des connaissances à partir de nombreuses ressources avec différents niveaux de détails) et se répercute ensuite durant la phase de stockage (généralement, plusieurs modèles de connaissances peuvent coexister dans un magasin de triples), et lors des deux prochaines étapes: les granules de connaissances utilisés lors de la diffusion et de l'intégration de connaissances peuvent être formulés à partir de différents points de vue et niveaux (dépendant de la diversité d'objectifs et d'utilisateurs considérée).

La granulation d'un contexte est une partition de l'observation qui conduit à une granulation des connaissances (Barg et al. 2000). Le critère de granulation est le contexte dans lequel le phénomène (problème, concept) peut se produire. Différemment de ce qui est généralement supposé dans le domaine de la granulation (Zadeh 1979) (Pawlak 1998), ce type de granularité ne doit pas nécessairement avoir une structure hiérarchique, mais le plus souvent des structures qui se chevauchent. Cela signifie que l'observation unique peut être vue à travers le prisme de plusieurs contextes simultanément ou à travers un contexte unique.

Dès lors, et dans mes recherches en cours et futures, je souhaite investiguer l'intégration d'approches de granulation avec des systèmes à base d'ontologies et les différentes problématiques sous-jacentes. En partant de l'étude de théories de partitions granulaires, comme celle de Bittner et Smith (Bittner & Smith 2001) ou encore l'approche plus récente de (Calegari & Ciucci 2010), je souhaite pouvoir identifier quelle approche est mieux adaptée au contexte d'application considéré. Chaque contexte (utilisateur ou d'utilisation) induit des besoins et des contraintes spécifiques et différentes, par exemple le besoin de s'abstraire de la situation ou du domaine à représenter, le besoin de comparer des abstractions différentes d'une même situation, ou encore l'incomplétude inhérente à la connaissance. Dans ce qui suit, je vais présenter comment cette problématique se retrouve dans mes activités de recherche en cours et l'impact qu'elle aura sur mes recherches prospectives à moyen terme.

3.4.1 Informatique granulaire et points de vue

Des considérations précédentes, le fait de diviser une représentation d'un domaine de connaissance en plusieurs modules plus petits semble être une bonne idée, favorisant la manipulation de la représentation générale et dès lors son implémentation. Par contre, cela n'apporte pas que des avantages: une telle représentation est souvent plus complexe à concevoir, et pouvoir manipuler ces points de vue de manière indépendante, tout en permettant leur interprétation conformément au modèle global, soulève des défis considérables.

Afin de mieux les illustrer, je vais présenter dans quel contexte je m'y attaque, à savoir dans le cadre du co-encadrement des recherches de Madame Claire PRUDHOMME. Débutée au milieu de l'année 2016, cette thèse se déroule en collaboration avec l'institut i3Mainz à Mayence, en Allemagne et est co-dirigée par mon collègue MCF HDR Christophe Cruz et le Prof. Franck Boochs (i3Mainz). La doctorante est d'ailleurs basée à Mayence, et des réunions mensuelles nous permettent de suivre ses avancements. Les travaux concernés par cette thèse visent la conception d'une approche pour maintenir une compréhension raisonnable à travers les interactions multi-agents et son implémentation dans un système d'aide à la décision pour la gestion de catastrophes naturelles (e.g. inondation). Dans ce contexte, un partenariat a été signé avec la ville de Cologne en Allemagne et l'institut i3Mainz. L'étudiante a ainsi pu utiliser de vraies procédures telles que mises en place par la ville de Cologne pour répondre aux différentes crues du Rhin.

La gestion de catastrophe (ou *disaster management*) n'est pas un domaine nouveau de recherche, aussi bien à un niveau international qu'à mon niveau. En effet, je m'y suis déjà intéressée de par le passé, notamment dans le cadre du projet régional PM2G (Plateforme Multi-Métiers Géolocalisés) dont le but était de permettre la coordination et la collaboration entre pompiers et agents de la DDE pour la gestion du trafic routier lors d'épisodes neigeux importants. Déjà à l'époque, les technologies sémantiques avaient aidé à spécifier la connaissance propre aux différents corps de métier, de même que l'ontologie de la viabilité hivernale conçue avait aidé à structurer les différentes procédures mises en place par la DDE. Toutefois, étant donné que presqu'une décennie s'est écoulée entre la fin du projet PM2G et le début de la thèse de Madame PRUDHOMME, un état de l'art solide du domaine a été établi. Cela nous a permis, d'une part, d'identifier la phase exacte de la gestion de catastrophes qui sera impactée par nos recherches, et, d'autre part, de définir les orientations recherches qui allaient être poursuivies. Les paragraphes suivants en donnent un bref résumé (la description des recherches poursuivies dans ce contexte sera plus courte que pour les travaux précédemment présentés, d'une part car je n'interviens qu'en tant que simple encadrante et dans une moindre mesure - seulement 30%, et d'autre part, les travaux présentés ici n'ont pas encore été finalisés et implémentés).

La gestion de catastrophes comprend 4 phases distinctes :

- La prévention concerne les activités mises en place afin de se protéger des catastrophes (e.g. plans d'évacuation).
- La préparation concerne les activités visant à limiter la perte de vies et les dommages matériels (e.g. déplacer des populations d'une zone à risque). Les actions de préparation représentent le principal levier permettant de réduire l'impact d'une catastrophe.
- La réponse comprend les actions coordonnées effectuées au moment où la catastrophe survient, ainsi que leurs effets à long terme. Ce type d'activité inclut le sauvetage, la fourniture d'eau et de nourriture, la prévention des maladies, la réparation des infrastructures vitales (e.g. telecommunications, transport)
- La récupération se produit une fois que la catastrophe est passée, mais avant que les personnes et les structures affectées soient totalement hors de danger. Les activités de récupération concernent donc la reconstruction d'infrastructures, la réhabilitation et la fourniture de soins.

Depuis janvier 2005, 168 pays ont adopté la déclaration de Hygogo, définissant un plan d'action sur 10 ans, en vue d'augmenter la résilience d'un pays face aux catastrophes "provoquées par des aléas d'origine naturelle ou imputables à des aléas ou risques environnementaux et technologiques connexes". Depuis 2015, l'ensemble des pays membres de l'ONU ont adopté le cadre d'action de Sendai pour la réduction des risques de catastrophes. En vigueur jusqu'en 2030, ce plan préconise 4 actions prioritaires, à savoir (Aitsi-selmi et al. 2015):

- Comprendre les risques de catastrophes, et ce sous toutes leurs dimensions: "la vulnérabilité, les capacités et l'exposition des personnes et des biens, les caractéristiques des aléas et l'environnement. "
- Renforcer la gouvernance des risques de catastrophe pour mieux les gérer: cela "suppose d'avoir une vision claire des choses, des plans, des compétences et des orientations, de coordonner l'action de tous les secteurs et d'un secteur à l'autre, et de faire participer toutes les parties prenantes."
- Investir dans la réduction des risques de catastrophe à des fins de résilience: mesures (non-)structurelles pour renforcer la résilience économique, sociale, sanitaire et culturelle des personnes
- Renforcer l'état de préparation aux catastrophes pour intervenir de manière efficace et pour "mieux reconstruire" durant la phase de relèvement, de remise en état et de reconstruction.

C'est au niveau du dernier point que se situent nos contributions dans ce domaine. Les recherches ménées concernent l'évaluation des plans d'évacuation et d'autres mesures, afin d'améliorer la collaboration entre les différents acteurs impliqués dans l'exécution du plan ou de la mesure, et ce au moment où la catastrophe survient (phase réponse). Pour ce faire, notre approche fut d'utiliser des simulations à base de systèmes multi-agents. Ce choix a une double justification:

- tout d'abord, les études montrent que ces simulations sont largement utilisées pour identifier une stratégie optimale ou des problèmes potentiels (Hawe et al. 2012)
- ce fut une direction donnée par les directeurs de la thèse que d'examiner les possibilités fournies par les systèmes à base d'agents dans l'approximation de processus humains

Notre approche de recherche propose de combiner ***un système d'information intégrant des données hétérogènes spatiales et un système de simulation multi-agent*** pour évaluer les différentes stratégies de coordination et de collaboration, telles que suivies par les agents. La question sous-jacente est de ***déterminer comment les agents partagent, utilisent et disséminent les informations dont ils disposent, tout en gérant l'hétérogénéité de ces informations et les problèmes associés en termes d'interopérabilité, de droits d'accès ou encore de qualité***.

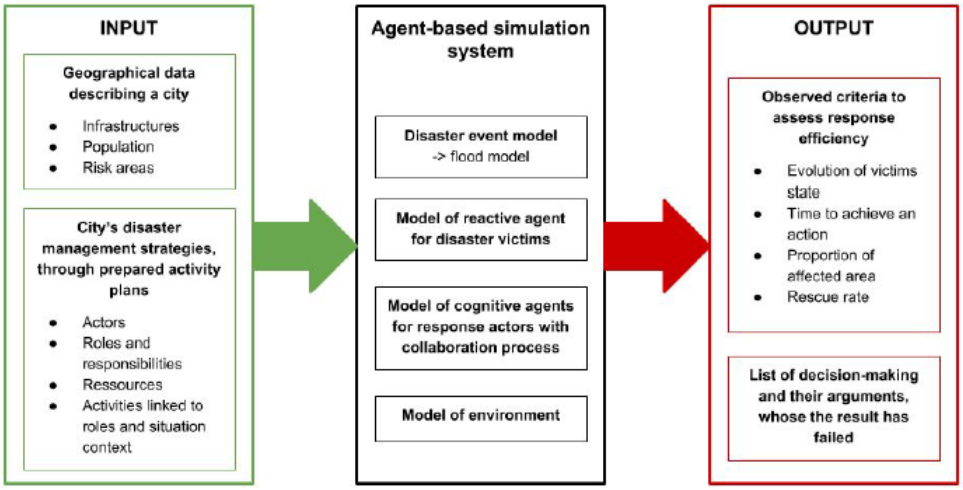


Figure 65. Schéma de notre approche pour le système de simulation à base d'agents.

De par mes recherches passées, j'ai suggéré d'utiliser une approche à base de technologies sémantiques et ainsi passer d'informations à connaissances. D'après une étude détaillée, le choix a finalement été fait d'utiliser une base de connaissance pour relier le système d'information et le système de simulation. Cette base de connaissance a un double intérêt: d'une part elle contient les connaissances nécessaires aux agents, d'autre part elle supporte les tâches de raisonnement des différents agents. La collaboration et la coordination entre agents permettront la création dynamique de simulations (à travers la spécification des différents types d'agents, leur modèle de comportement et le modèle de catastrophe considéré).

L'idée de permettre à des systèmes de simulation multi-agents d'implémenter des capacités de raisonnement symbolique ne date pas d'aujourd'hui. Initialement, les recherches en intelligence artificielle ont étudié comment un seul agent peut être doté d'une intelligence propre (singulière et interne à l'agent).

Cependant, ces dernières années, nous avons observé un intérêt pour la concurrence et la distribution dans l'IA, qui ont donné naissance à la branche appelée intelligence artificielle répartie (*Distributed Artificial Intelligence* ou DAI). Cette discipline relativement récente comprend deux domaines principaux: la résolution répartie de problèmes (*Distributed Problem Solving* ou DPS) et les systèmes multi-agents (*Multi-Agent Systems* ou MAS). Plusieurs exemples d'applications utilisant des approches multi-agents pour l'aide à la décision sont présents dans la littérature, que ce soit pour la découverte de modèles de parieurs en bourse (Bac et al. 2012; Korczak et al. 2013), pour l'analyse de journaux de navigation Web (Weichbroth & Owoc 2011) ou encore pour l'évaluation de technologies de l'information (Orłowski et al. 2010).

Permettre à des agents d'un MAS de raisonner sur des connaissances implique les verrous suivants: la coordination des connaissances, des objectifs et des actions parmi les agents autonomes, la synthèse des connaissances au niveau agent, le raisonnement en utilisant des règles logiques appliqués selon un contexte.

Lorsqu'il s'agit d'utiliser un raisonnement symbolique afin de fournir un comportement intelligent à des agents dans un système de simulation multi-agents, une approche courante est d'utiliser des agents à raisonnement déductif (*Deductive Reasoning Agents*). Ces agents utilisent une théorie logique pour définir les actions à effectuer dans une situation donnée. Il s'agit du type d'agent utilisant une spécification formelle de leurs connaissances. Malheureusement, ils souffrent des limitations associées à ces spécifications formelles: premièrement, la complexité des preuves de théorèmes (pouvant même mener à des situations indécidables) et deuxièmement, les limites d'expressivité dues à l'hypothèse du monde ouvert (et de la monotonie de la connaissance).

Le raisonnement déductif nécessite, quant à lui, de sélectionner des logiques sous-jacentes prennant en charge la nature des agents. Il convient de mentionner que les implémentations les plus importantes des agents à raisonnement déductif sont basées sur des logiques intentionnelles comme les modèles formels de logique intentionnelle (Rao & George 1995) (e.g. *Belief - Desire - Intention* ou BDI), qui prennent en compte un sous-ensemble de la logique modale de Saul Kripke (Kripke 1980).

Les problèmes de raisonnement symbolique ont conduit à la création du «mouvement des agents réactifs» en 1985, marquant le début d'une période où les recherches se focaliseront sur les architectures à base d'agents réactifs. Ce mouvement a donné naissance à un ensemble d'exigences pour les langages dits de comportement (*behaviour languages*) (Brooks 1991):

1. Un comportement intelligent peut être généré sans représentations explicites comme en IA symbolique.
2. Un comportement intelligent peut être généré sans raisonnement abstrait explicite comme en IA symbolique.
3. L'intelligence est une propriété émergente de certains systèmes complexes.

De nos jours, les approches à base d'agents réactifs sont communes (Kaplanski & Weichbroth 2017), mais manquent, la plupart du temps, de fondements formels et ce type de système de simulation multi-agents est très difficile à analyser avec des méthodes et des outils formels. Néanmoins, le mouvement des agents réactifs a permis le développement de la programmation orientée agent (AOP), e.g. JADE (Bellifemine et al. 1999). Notre étude de l'état de l'art dans le domaine des systèmes de simulation multi-agents a été publiée en 2017 [CI5].

Sur la base de cette étude, dans le contexte de notre approche, nous avons choisi de nous baser sur le protocole ODD (*Overview, Design concepts, Details*) (Bouquet et al. 2015), reconnu par la communauté en tant que référence pour la description de modèles de simulation. Nous avons souhaité déterminer la faisabilité d'une approche basée sur une architecture hybride, combinant les architectures symboliques et réactives. Notre système comprend deux sous-systèmes: l'environnement symbolique (supportant le processus de prise de décision des agents) et un moteur de raisonnement réactif (réagissant à des événements précis, sur la base de règles logiques, et déclenchant des raisonnements simples). Pour les différents types d'agents considérés, nous nous sommes inspirés de l'architecture d'agents utilisée par la machine de Touring, telle que présentée par Ferguson dans (Ferguson 1992): "*The TouringMachine agent architecture comprises three separate control layers: a reactive layer, a planning layer, and a modelling layer. The three layers are concurrently-operating, independently-motivated, and activity-producing: not only*

is each one independently connected to the agent's sensory apparatus and has its own internal computational mechanisms for processing appropriate aspects of the received perceptual information, but they are also individually connected to the agent's effectory apparatus to which they send, when required, appropriate motor-control and communicative action commands".

A travers le co-encadrement de Madame PRUDHOMME, j'ai pu renouer avec une thématique de recherche que je poursuis depuis la soutenance de ma thèse de doctorat, à savoir le Web sémantique spatial. En effet, dès 2009 je me suis intéressée aux apports des technologies sémantiques pour la gestion de la qualité des informations géographiques e.g. leur précision, degré de mise à jour, exhaustivité, cohérence logique, etc. J'ai investigué la construction d'ontologies spatio-temporelles pour la modélisation de contenus nécessitant une localisation dans l'espace et dans le temps (modélisation d'informations géographiques et services à base de localisation, étude des standards ISO et OpenGIS). Ces recherches ont été appliquées dans le cadre du projet PM2G (ontologie de la viabilité hivernale dans le Territoire de Belfort) [CH9] et au travers de ma collaboration avec le Laboratoire de Topométrie de l'EPFL (description sémantique de service Web permettant de traduire une position géographique, en passant d'un système de repérage à un autre).

Sur la base de mes recherches passées, il y a trois principales questions recherche qui m'intéressent plus particulièrement:

> ***Quelles informations doivent être partagées entre agents pour satisfaire un but donné*** ? Nous retrouvons ici la discussion précédente sur les apports de la granularité dans l'intégration de différents points de vue. En effet, dans notre approche, nous considérons deux types d'agents: les agents réactifs (en charge de sauver des victimes humaines) et les agents cognitifs (représentant des acteurs du monde réel e.g. pompiers et collaborant avec d'autres agents cognitifs). La principale différence entre les deux est que l'agent réactif ne correspond pas à une entité réelle. Un agent réactif doit "simplement" percevoir son environnement et réagir à cette perception. Cela a l'air simple, mais en réalité plusieurs questions se posent: faut-il limiter la perception qu'a un tel agent de son environnement ? si oui, par rapport à quoi et comment ? Cela revient à spécifier avec quel niveau de granularité l'agent réactif doit percevoir le monde réel afin de permettre une prise de décision rapide. En effet, plus la perception de l'agent réactif est complexe, plus le raisonnement sur cette perception sera long et difficile à implémenter [CI5]. Cette question est aussi pertinente pour les agents cognitifs, qui eux possèdent ce qu'on appelle une représentation du monde qui se traduit sous la forme d'un ensemble de croyances déterminant les actions à entreprendre, toujours par rapport à une perception (Woolridge & Wooldridge 2001). Il est dès lors tout à fait pertinent de s'interroger sur la notion de granule d'information dans ce contexte. Définir des granularités au niveau des données manipulées par les agents mènera à la spécification de niveaux de granularité de l'information nécessaires pour supporter une prise de décision efficace.
> ***Quel est l'impact a la qualité de données utilisées sur les résultats de l'évaluation*** ? Les principales données considérées en entrée du système sont des données géospatiales et des données spécifiant les différents acteurs intervenant, leurs rôles et les actions à entreprendre dans les différents contextes, de même que les procédures d'évaluation des risques d'inondation telles qu'utilisées par la ville de Cologne. Autant les données fournies par la ville ont pu être intégrées au sein d'une ontologie modélisant les procédures de gestion d'inondations par la ville de Cologne, autant pour l'intégration des données spatiales cela ne fut pas aussi rapide. Les données géospatiales concernent des infrastructures (e.g. ponts, routes), la topographie du terrain, la répartition de la population ou encore les courants d'eau de la municipalité. Une des premières questions auxquelles nous avons dû répondre fut de déterminer de quoi traite une donnée géospatiale afin de savoir dans quel processus de prise de décision elle peut être pertinente. En ce sens, nous avons proposé une nouvelle approche exploitant les technologies du Web sémantique pour une interprétation de ces données [CI6]. Il s'agit maintenant d'identifier si l'interprétation des données géospatiales ne pourrait être couplée aux niveaux de granularité de données restants à définir, afin de relier automatiquement une qualité de données géospatiales à un niveau de perception requis pour un agent cognitif.
> ***Comment permettre l'interopérabilité entre les différents modèles utilisés par le système*** ? Pour chaque type de comportement pouvant être implémenté par un agent, un sous-modèle de comportement est défini.. Ce rôle central implémente deux comportements: d'abord identifier une correspondance entre les besoins et les possibilités (grosso modo détermine la tâche à réaliser selon les informations fournies en entrée), puis la coordination proprement dite visant plus spécifiquement le partage de ressources

entre plusieurs acteurs. Vient ensuite le modèle du gestionnaire, lui-même pouvant implémenter deux comportements: 1) découpage d'une tâche complexe en sous-tâches, détermination des informations nécessaires à chacune d'entre elles et attribution d'une sous-tâche à un agent acteur, 2) planifier de nouvelles tâches pour atteindre un but raté de par le passé. Enfin, l'agent acteur revêt le rôle opérationnel (applique une tâche). La tâche d'un agent acteur est, du point de vue de l'agent gestionnaire, comme un ensemble d'actions. Ces modèles n'ont pas encore été formalisés, la question de leur fédération demeure ouverte. Je veillerai à réutiliser l'architecture fédérée FOWLA si toutefois cela est possible avec le formalisme retenu pour la représentation de ces modèles.

Jusqu'ici, le rayonnement scientifique associé aux travaux de recherche présentées ici a donné lieu à:

> 1 article publié dans une Revue Internationale Sélective à Comité de Lecture [RIS4]
> 4 communications à des congrès internationaux à comité de sélection et actes publiés [CI2], [CI5], [CI6], [CI9]
> 2 chapitres d'ouvrages [CH2], [CH3]

Les recherches présentées ici sont encore en cours. La doctorante doit terminer de formaliser le modèle de représentation des agents ainsi que le moteur réactif à base de règles SWRL. Par la suite, ces composants devront être intégrés dans une application type preuve de concept, afin de pouvoir évaluer l'approche en termes de précision et d'efficacité. En parallèle, la doctorante a commencé la rédaction de son manuscrit de thèse (en anglais, afin de pouvoir solliciter le label européen pour sa thèse de doctorat).

3.4.2 Informatique granulaire et capteurs

Suite au dépôt de la réponse à appels à projets H2020, BigSTEP (Big Semantic spatio-Temporal data for built Environment Processes) dont j'ai été la coordinatrice, un sujet de thèse a été rédigé et proposé comme candidat pour une bourse de l'Ecole Doctorale SPIM. Le projet de thèse spécifiait un encadrement des travaux de recherche par Prof. Dominique GINHAC (directeur LE2I), MCF HDR Christophe CRUZ et moi-même. Le financement de la thèse associée a été accepté par l'Ecole Doctorale, et en février 2017, nous avons recruté Monsieur Muhammad ARSLAN pour effectuer sa thèse de doctorat sur le sujet.

Les recherches à mener se placent dans le contexte de l'environnement bâti au sens large (allant au-délà de la limite d'un seul bâtiment) et tente d'apporter une réponse à des problèmes identifiés au travers du projet BigSTEP ainsi que de par mes activités au sein d'organisations telles bSI. En effet, les acteurs de l'environnement bâti (concepteurs, gestionnaires techniques de patrimoine) font souvent face à des problèmes liés à l'accès et au partage de données, d'une part, et à l'intégration de volumes massifs de données provenant de sources hétérogènes, d'autre part. Ces limites impactent directement le niveau d'efficacité du traitement et de l'analyse de ces données, et représentent des problèmes communs dans les domaines de l'architecture, l'ingénierie et la construction (AEC). Ainsi, les professionnels de l'environnement bâti ne s'appuient pas seulement sur des données relatives aux éléments de construction. Les bâtiments ne sont pas des systèmes isolés, mais ils évoluent dans un écosystème beaucoup plus large - l'écosystème urbain. Les capteurs intelligents sont aujourd'hui parties intégrantes (e.g. capteurs de température, qualité de l'air, capteur de luminosité, etc.). Cet écosystème implique un nombre grandissant de services et d'applications. Même en disposant d'accès à des données standardisées sur le bâti (e.g. fichiers IFC), l'intégration de différentes sources de données donne des avantages concurrentiels considérables aux acteurs du secteur.

Par conséquent, une approche uniforme est requise pour modéliser les données, ainsi que les processus impliquant les différents niveaux de granularité, par exemple du niveau de la surveillance de capteurs temps-réel au niveau de la planification urbaine. L'objectif pluridisciplinaire qu'ambitionne ce projet est de proposer une infrastructure intelligente de traitement de données sémantiques, spatio-temporelles et massives, provenant des capteurs afin de fournir des services intelligents dans le cadre de la gestion et des usages des bâtiments.

Les travaux de cette thèse s'orienteront sur deux aspects fondamentaux consistant à:

a) définir une architecture de traitement de données massives provenant de capteurs intelligents sources de données quantitatives et qualitatives.
b) piloter en temps réel des capteurs en fonction de l'évolution de l'environnement du bâti. Le pilotage des capteurs consiste soit à faire évoluer la manière dont sont captées les données, soit la coordination des capteurs afin de faire remonter une information spécifique.

La modélisation des connaissances, fondation de l'architecture, permettra de définir les aspects symboliques traités par le système intelligent. La sémantique ainsi modélisée offrira un support de qualification des informations, et d'inférence sur les connaissances. Ce support, dans le cadre d'une architecture Big Data, devrait permettre une adaptation dynamique des capteurs aux besoins de gestion et d'usages du bâti.

A travers l'encadrement des recherches de Monsieur ARSLAN, j'ai pu renouer avec une thématique de recherche que je poursuis depuis la soutenance de ma thèse de doctorat, à savoir le Web sémantique spatial. Toutefois, ici le verrou recherche ne concerne pas particulièrement la qualité des données. Même si elle est souvent approximée par les concepts de "valeur" et "véracité" des données Big Data, dans le contexte de cette thèse nous allons certainement ré-utiliser l'approche définie dans [CI6] pour évaluer la qualité des données utilisées. L'accent sera plutôt mis sur des problématiques de recherche différentes, concernant la définition de relations spatiales (e.g. relations topologiques) et de concepts spatiaux à l'aide d'axiomes contenant des prédicats spatiaux, ainsi que la possibilité de raisonner sur la spatialité des instances (pouvoir inférer des relations spatiales). En ce sens, je m'intéresse à la définition d'une théorie de partitions granulaires ou à l'utilisation d'une approche existante.

Le traitement automatique des données et l'extraction de règles à partir de grands ensembles de données ont suscité un intérêt considérable au cours des dernières années. Plusieurs approches ont été proposées,

basées à la fois sur des statistiques et de la logique floue. Dès lors que l'on souhaite utiliser les concepts de l'informatique granulaire, les contributions se font plus rares (Bittner & Smith 2003), (Grenon & Smith 2004). Or, l'intérêt récemment renouvelé (Pedrysz 2007) et croissant de la communauté scientifique pour ses idées et principes a prouvé peuvent s'avérer très utiles dès lors qu'il s'agit d'intégrer des sources hétérogènes de données Big Data. Ainsi, dans (Xu & Yu 2017), les auteurs présentent une approche exploitant les granules d'information pour la sélection de sources de données de confiance ainsi que pour la définition d'une nouvelle pour la fusion d'informations. Dans (Butenkov et al. 2017), les auteurs reprennent les méthodes de granulation de l'information définies par (Zadeh 1979) et les étendent afin de définir une interprétation géométrique d'un espace de données supportant l'implémentation d'analyses intelligentes de données à multi-dimensions (Rogozov et al. 2013). Dans ce contexte, pour l'instant un seul état de l'art a été réalisé discutant des approches existantes utilisant des données Big Data pour la gestion de catastrophes [CI4]. Nous souhaitons étudier davantage les contributions définies au travers de projets européens tels les projets européens City Pulse (EU FP7), Ready4SmartCities (EU FP7) ou encore le projet DURAARK (EU FP6).

Dès lors que l'on s'intéresse aux aspects spatio-temporels des données et leur intégration dans des ontologies, les différentes approches ont souligné la nécessité de compléter les représentations méréologiques par des concepts et des principes topologiques. Ceci peut se justifier à travers deux principales raisons. La première raison est liée à la caractérisation de l'intégrité d'un individu ou encore de l'unité organique: étant donné que spécifier une connexion entre parties est contraire à la méréologie simple, une théorie des parties et des ensembles doit intégrer un mécanisme topologique quelconque. La seconde raison a été mise en avant principalement par les besoins de raisonnement sur la spatialité et la temporalité des instances: dans ce contexte, la méréologie fournit des mécanismes simples pour la définition de liens entre choses ou événements; toutefois, il faut définir une topologie afin de tenir compte du fait que deux événements peuvent être continus l'un avec l'autre, ou pour spécifier des relations telles à l'intérieur de, à l'extérieur de, en butée de ou autour de quelque chose d'autre. En ce sens, je souhaite exploiter les éléments présentés dans (Varzi 1993), notamment les discussions sur les trois stratégies permettant de coupler méréologie et topologie:

> ***Première approche*** - ajouter la topologie au dessus d'une base méréologique. Dans ce contexte, la méréologie représente la théorie de base sur laquelle sont construites des théories de plus en plus complexes (y compris la topologie, la morphologie ou encore la cinématique), en spécifiant les notions et principes nécessaires.
> ***Deuxième approche*** - partir de la topologie et définir des notions méréologiques en utilisant des primitives topologiques. Comme la méréologie peut être vue en tant qu'une généralisation d'un théorème de l'identité (encore plus fondamental), la topologie peut être intérprétée comme une généralisation de la méréologie (les relations de jointure ou de connection deviennent des adaptations spécifiques des relations de chevauchement ou "faire partie de").
> ***Troisième approche*** - revendication de la méréologie exploitant sa généralité formelle. Dans ce cas, la topologie est une partie de la méréologie, spécifique à un domaine. Des relations comme la connection (et les relations apparentées) sont traitées sous la forme de relations du type "fait partie de" entre entités d'un certain type.

Pour l'instant, les travaux menés dans ce contexte ont investigué les approches de représentation de trajectoires spatio-temporelles avec des technologies sémantiques. Le sujet n'est pas nouveau, et l'approche de (Parent et al. 2008) reste encore une des mieux construites et spécifiées. Selon les auteurs de (Parent et al. 2008), une trajectoire sémantique représente une interprétation qualitative d'un mouvement, tout en préservant les aspects temporels (une trajectoire est divisée en plusieurs périodes de temps). Avec cette approche, une trajectoire sémantique est représentée par le biais de ses attributs topologiques et sémantiques, favorisant dès lors l'adressage de requêtes SPARQL. Dans ce contexte, après une étude des approches existantes (fournie par l'étudiant sous la forme d'un rapport technique), nous avons publié un article de journal à la fin de l'année dernière [RIS2].

A terme, je souhaite profiter de ce projet de recherche pour étudier plus en détail le domaine des capteurs. En ce sens, je souhaite analyser les algorithmes d'apprentissage par renforcement et leurs avantages éventuels lorsqu'utilisés dans une architecture de contrôle de capteurs. Selon (Girard & Khamassi 2016), deux principaux types d'algorithmes existent dans la littérature: ceux utilisant un modèle interne de l'environnement (pour prendre des décisions) et ceux n'en utilisant pas (apprenant par cœur

les bonnes associations (objectif, action)). Toujours d'après (Girard & Khamassi 2016), la combinaison de ces deux types d'algorithmes au sein d'une "même architecture de contrôle permet [...] d'envisager de tirer le meilleur parti des deux mondes". De plus, (Girard & Khamassi 2016) discutent et évaluent différentes méthodes de combinaison d'algorithmes. Toujours concernant les capteurs, je souhaite tirer parti de ma participation à des organismes de standardisation afin d'avoir une meilleure connaissance des niveaux de confiance et de crédibilité que les utilisateurs associent aux différents standard actuellement existant pour les échanges de données, la protocoles de communication ou encore les formats de données.

Un dernier point d'attention concerne l'implémentation d'un processus de raisonnement à partir de données capteurs. L'idée serait d'exploiter les fonctionnalités de calculs personnalisables qu'il est possible de retrouver au niveau des systèmes de contrôle commande et déterminer si elle peuvent être soit implémentées au niveau capteur (et concevoir ainsi des capteurs dits proactifs, réagissant sur la base de règles du type "SI condition ALORS action") ou alors exploiter des approches de décentralisation des traitements dans le capteur (à travers la définition et le couplage de blocs fonction comme par exemple dans la norme IEC 61804). A terme, l'idée est de concevoir une architecture pour capteurs dits intelligents (smart sensors). Cela suppose une structuration préalable des données, et notamment la définition de méthodes pour le traitement, la classification ou encore le stockage des (grandes quantités de) données produites. En ce sens, les données seront transformées en connaissances, par le biais de différents vocabulaires sémantiques, et une annotation sémantique du flux de données capteurs sera recherchée. Enfin, une attention particulière sera accordée au système de contrôle commande utilisé pour le pilotage temps-réel des différents capteurs. Ce système étant constitué d'éléments hétérogènes, il faut analyser les différents risques informatiques et menaces et proposer une approche globale pour assurer sa cybersécurité.

Jusqu'ici, le rayonnement scientifique associé aux travaux de recherche présentées ici a donné lieu à:

> 1 article publié dans une Revue Internationale Sélective à Comité de Lecture [RIS2]
> 2 communications à des congrès internationaux à comité de sélection et actes publiés [CI3], [CI4]

Les recherches présentées ici sont à peine débutées. Par rapport aux considérations ci-dessus, il reste plusieurs états de l'art à faire (capteurs intelligents, algorithmes d'apprentissage, etc.). Ces études nous serviront de base pour les discussions définissant les approches à concevoir ou à adapter. La discussion concernant l'approche à utiliser pour approximer la trajectoire d'un individu dans un espace urbain n'est toujours pas close. Sur la base des différentes approches combinant méréologie et topologie, il s'agit de choisir celle qui convient le mieux pour représenter ce type de trajectoire (contenant des sous-parties et ayant une continuité dans le temps). De plus, plusieurs des problématiques exposées ci-dessus sont commune savec celle traitée dans le cadre du projet ANR McBIM (dont je suis la responsable scientifique pour le LE2I). Il s'agit de l'annotation sémantique de flux de données capteurs, leur intégration dans un SI (qu'il soit géographique ou pas) afin de pouvoir raisonner sur ces données et enfin la sécurisation des différents échanges de données.

3.5 Conclusions

3.5.1 Résumé des recherches effectuées

Les sections et chapitres précédents ont retracé et détaillé les différentes recherches menées et co-encadrées depuis ma nomination en tant que Maître de Conférences en 2012. Les paragraphes ci-dessous les résument par rapport aux axes de recherche définis en introduction du chapitre 3.

- **Axe 1 – L'utilisateur humain – Sensibilité au contexte de l'utilisateur**
  - Dans le cadre du co-encadrement de la thèse de Madame Anett HOPPE, j'ai travaillé à la définition d'un profilage utilisateur plus fin (que fait précédemment durant ma thèse). Nous avons défini une méthode permettant le peuplement automatique d'une ontologie de profil utilisateur, et ce avec des informations obtenues à partir de cookies. Nous avons ensuite mis au point une méthode (à base de règles logiques) permettant d'associer un profil utilisateur à un segment marketing, et ce avec une certaine probabilité (calculée à partir de données extraites des cookies et en utilisant des règles logiques SWRL). Dans ces travaux, nous n'avons pas utilisé de données liées – les ontologies et les règles utilisées ont été développées manuellement, à partir d'échanges avec des experts en marketing et en publicité en ligne. L'implémentation sous forme de preuve de concept de ces approches, nous a permis d'identifier des limites claires en ce qui concerne les possibilités de raisonnement offertes. Du fait du niveau d'expressivité des ontologies et de la complexité des règles logiques, le calcul des pondérations pour l'appariement du profil utilisateurs aux différents segments marketing était particulièrement long. C'est sur la base de ces enseignements que, par la suite, j'ai cherché à trouver le meilleur compris entre niveau d'expressivité du modèle et rapidité d'exécution des requêtes. Ceci se retrouve dans mes travaux concernant les 2 autres axes.
- **Axe 2 – Contexte proche – Interopérabilité sémantique des Systèmes d'Information**
  - Au travers du co-encadrement de la thèse de Monsieur Tarcisio MENDES de FARIAS, j'ai pu appliquer les travaux effectués dans cet axe au domaine de l'architecture et la construction, plus particulièrement dans le domaine du Building Information Modeling (BIM). En effet, dans ce domaine il existe plusieurs standards pour la description des données bâtiment, et il y a une réelle problématique par rapport à l'interopérabilité entre ces modèles, surtout dans un contexte de passage au numérique des processus humains sous-jacents (e.g maquette numérique, gestion de patrimoine). J'ai ainsi cherché à appliquer les principes des données liées à l'adaptation en ontologie du standard ISO 16739 :2013 ou IFC (Classe de fondation d'industrie) pour le partage des données dans le secteur de la construction et de la gestion des installations. Nous avons proposé une nouvelle approche, ifcWoD (Web of Data) qui permet de réduire la longueur des requêtes SPARQL mais aussi leur temps d'exécution, tout en retournant le bon nombre de résultats. Toujours pour adresser le problème d'extraction d'informations d'un fichier IFC, nous avons développé un algorithme qui permet d'extraire des éléments d'une ontologie ifcOWL (en mentionnant soit le nom de concept ou de relation, soit son identifiant IFC) puis de générer un fichier IFC au format EXPRESS (fichier pouvant être manipulé avec des logiciels du commerce). Par la suite, avec le thésard, nous avons défini une ontologie OWL de la norme internationale COBie (Construction-Operations Building information exchange). Les données COBie, définies dans des fichiers *.csv ou *.xls, définissent des informations pour les actifs livrés dans le cadre d'un projet de construction d'installation. La norme COBie est fortement utilisée pour documenter les données projets BIM, car considérée plus intuitive et plus facile à comprendre que le standard IFC. Une fois adaptée en ontologie (COBieOWL), nous avons défini un alignement (à travers 474 règles logiques SWRL) vers le standard IFC. Notre approche permet donc d'utiliser des données COBie pour peupler un fichier IFC et vice versa. Toutefois, suite à l'implémentation de cet alignement à base de règles entre les deux ontologies (ifcWoD et COBieOWL), nous nous sommes aperçus qu'il était très difficile de le requêter. En effet, avec chaque requête adressée, l'ensemble des règles définies étaient aussi exécutées, ralentissant l'ensemble du processus, et même provoquant des situations où il nous fallait davantage de mémoire pour pouvoir traiter la requête (out of memory). Nous avons donc réfléchi et proposé une nouvelle approche (FOWLA) qui permet de sélectionner uniquement les règles logiques pertinentes par rapport à une requête

donnée. Cette approche a été formalisée en dehors du domaine du BIM et peut être adaptée pour requêter de manière efficace tout alignement d'ontologies à base de règles.

- **Axe 3 – Contexte étendu – Web sémantique spatial**
  - C'est au travers du projet semanticGIS et notamment à travers l'encadrement des travaux de Madame Claire PRUDHOMME, et plus récemment Monsieur Muhammad ARSLAN qui j'ai fait avancer mes recherches dans ce domaine. Dans le cadre de la thèse de Madame PRUDHOMME des recherches ont été menées sur l'intégration et la qualification de données géographiques hétérogènes, la modélisation de processus humains avec des systèmes multi-agents et enfin l'utilisation d'une ontologie de domaine pour pouvoir tester l'orchestration de ces processus et ainsi pouvoir évaluer des plans de gestion de catastrophe (application aux inondations et collaboration avec la ville de Cologne en Allemagne). Grâce au co-encadrement de la thèse de Monsieur ARLSAN, j'ai pu reprendre mes travaux sur la modélisation et l'interopérabilité d'informations géographiques avec des technologies sémantiques. Il s'agit de s'attaquer aux problématiques des trajectoires des utilisateurs dans un contexte BIM/SIG, avec des questions sous-jacentes telles "comment caractériser une trajectoire à partir de points ?" (ce qui rejoint les recherches sur la spatialité des instances menées après ma thèse).

Pour conclure quant à l'équilibre à trouver entre expressivité des représentations de la connaissance et les possibilités de raisonnement sous-jacentes, je peux dire que cette question peut difficilement être résolue avec une approche globale, et nécessite des adaptations spécifiques. Nous venons de voir que lorsque l'on souhaite atteindre un haut niveau d'automatisme dans la manipulation de connaissances ontologiques (et donc aller vers un haut niveau d' "intelligence"), un haut niveau de formalité est requis. En effet, plus l'ontologie est complexe et "lourde", plus son niveau de formalisme est élevé et donc les processus pouvant être implémentés au dessus seront de plus en plus automatisés et puissants (Grimm et al. 2011). Bien évidemment, un haut degré de formalisme ne présente pas que des avantages. Quand il s'agit d'implémenter des approches à base d'ontologies, souvent cela revient à trouver le juste milieu entre coûts de conception/maintenance et niveau d'automatisme des processus. De plus, il existe des scénarios d'application où un haut niveau de formalisme n'est tout simplement pas nécessaire (notamment dans le cas de structurations informelles de la connaissance e.g. glossaires, qui sont conçues à destination d'utilisateurs humains).

Plus les connaissances modélisées dans une ontologie sont explicites, plus l'ontologie sera comprise et réutilisée par les humains, et pourra servir de base pour un appariement (semi-)automatique entre plusieurs ontologies traitant du même domaine de connaissance (Grimm et al. 2011). Afin d'être le plus explicite dans sa modélisation, il convient de rendre les choix de modélisation explicites, les modéliser de manière formelle ou encore modéliser des aspects redondants de la connaissance dans un langage expressif.

Si l'on vise l'interopérabilité entre données et modèles, il est important d'atteindre un consensus au niveau des choix de modélisation (Grimm et al. 2011). Il est important de savoir que pour un domaine de connaissance donnée, il y aura autant de modèles que de concepteurs. Le consensus entre modélisations semble dès lors difficile à atteindre entre ontologistes. Dans un premier temps, l'utilisation des standards du Web sémantique pour la description d'ontologies est un premier consensus sur les formalismes de représentation des connaissances (intéroperabilité schématique). J'ai démontré, dans la section 3.3.4.2.4, que l'appariement d'ontologies à base de règles logiques est une approche vers l'interopérabilité sémantique de tels modèles.

3.5.2 Projets de recherche prospectifs à moyen terme

Par rapport à ces observations, mon projet de recherche à moyen terme concerne l'informatique granulaire. Comme discuté dans la section 3.4, il existe encore un énorme fossé entre la résolution de problèmes avec des ordinateurs et la résolution de problèmes par les humains. La granularité est une nouvelle approche de structuration des représentations de la connaissance, orthogonale aux méthodes existantes de modélisation. Elle permet un découpage du domaine de connaissance ciblé selon des méthodes structurées et cohérentes, impliquant une définition des types de granularité considérés ainsi que des critères formant une perspective granulaire. Ceci peut être implémenté à différents niveaux de détail, de manière à éviter soit des enrichissements trop détaillés, soit des généralisations trop grossières.

Une telle approche permet de raisonner sur l'ensemble du domaine de connaissance ainsi modélisé, ainsi que à travers des différents niveaux de granularité considérés. L'implémentation de niveaux de granularité dans les représentations de la connaissance permettrait de réduire le fossé entre les processus de résolution de problèmes utilisés par les humains et les approches machines. Les verrous identifiés et les éléments de réponse identifiés ont été discutés par rapport à deux projets de recherche en cours, notamment ceux de Madame PRUDHOMME (voir section 3.4.1) et Monsieur ARLSAN (voir scetion 3.4.2).

De plus, de par mon implication dans les organismes de standardisation mais aussi au travers de mes collaborations avec les industriels, je sais que les problèmes discutés auparavant ne sont pas entièrement résolus, et qu'il reste de nombreux verrous à traiter dans les années à venir et ce pour chacun des 3 axes identifiés. Les paragraphes ci-dessous résument ces verrous, et essayent de lister des actions permettant de les adresser.

- **Axe 1 – L'utilisateur humain – Sensibilité au contexte de l'utilisateur**
  - ***Verrous*** : Maintenant que j'ai pu améliorer mes connaissances dans le domaine du bâtiment et des SIG, il s'agit de les exploiter afin de permettre une contextualisation de l'utilisateur par rapport à un environnement BIM/SIG. Des recherches dans cette direction me permettraient de répondre à des questions concernant l'identification d'usages de bâtiments comme de zones urbaines, mais aussi des questions portant sur l'identification de motifs dans l'utilisation que les humains ont des bâtiments (pattern usage). L'idée c'est par la suite de pouvoir investiguer les éventuels apports des méthodes de recherche d'information à base de motifs (pattern mining) pour pouvoir adapter la conception des bâtiments aux usages ou encore pour pouvoir identifier des erreurs de conception.
  - ***Actions à mener*** : J'envisage de poursuivre ces recherches au travers de 2 vecteurs : le co-encadrement de la thèse de Monsieur ARSLAN, et la collaboration avec le CSTB (qui se retrouve dans la publication [CI1] et au terme de laquelle est envisagé un sujet de thèse en co-encadrement).
- **Axe 2 – Contexte proche – Interopérabilité sémantique des Systèmes d'Information**
  - ***Verrous*** : Il s'agit ici de poursuivre les travaux débutés dans le domaine du BIM, pour améliorer la manipulation de maquettes numériques, la contextualiser par rapport à un besoin métier, ainsi que par rapport à une phase du cycle de vie du bâtiment. Je souhaite investiguer l'extraction de vues IFC adaptées à un métier ou une phase donnée (e.g. vue conception). Pour la définition de telles vues, les travaux passés ont mis en évidence la pertinence des approches à base de règles logiques ; je souhaite donc utiliser ce type d'approche pour spécifier de telles vues. Toutefois, de nouveaux standards sont apparus pour définir de telles règles (e.g. SHACL ou SPIN). J'envisage d'évaluer leurs éventuels avantages par rapport aux règles écrites jusqu'ici en SWRL. Ces recherches pourront être implémentées au travers de preuves de concept pour la vérification de (non-)respect de réglementations diverses (e.g. respect de normes incendie).
  - ***Actions à mener*** : Dans ce contexte, j'ai été sollicitée à de nombreuses reprises par des acteurs du domaine. Le CSTB m'a soumis un projet de co-encadrement de thèse (visant la vérification semi-automatique de maquettes numériques) ; l'association QUALITEL/Cerqual m'a aussi demandé d'encadrer un sujet de postdoc (sur l'identification de clauses de la norme NF Habitat non-respectées par les maquettes numériques) ; un projet de co-encadrement de postdoc avec Monsieur NICOLLE, et en collaboration avec l'entreprise NOBATEK, a démarré début 2018 pour une durée de 24 mois (visant le développement d'une fédération d'ontologies à base de règles).
- **Axe 3 – Contexte étendu – Web sémantique spatial**
  - ***Verrous*** : A terme, cet axe rejoindra l'axe 2, principalement parce qu'il ne reste plus beaucoup de problématiques à traiter qui soient spécifiques à cet axe. Les verrous concernent principalement l'intégration dans un SIG sémantique d'informations issues du BIM et annotées sémantiquement. Jusqu'ici, mes travaux dans le domaine de la gestion de

catastrophes ont uniquement utilisé des données géographiques (obtenues à partir de SIG). Il s'agira à l'avenir d'investiguer l'intégration d'informations bâtiment dans ce type d'approche, en veillant à des problématiques comme la continuité de positionnement, le maintien de la cohérence des informations, l'alignement des différents systèmes de positionnement utilisés (notamment le problème de l'intégration dans un SIG d'informations concernant le sous-sol des bâtiments). Une intégration sans coutures d'informations BIM avec des informations SIG permettrait d'avoir une meilleure connaissance de l'environnement urbain au sens large en apportant, par exemple, des connaissances sur les impacts sur la structure d'un bâtiment générés par un tremblement de terre ou par une infiltration d'eau dans les fondations.

> ***Actions à mener*** : De tels travaux feront l'objet du projet de recherche ANR McBIM, dans le cadre duquel je serais responsable de l'encadrement d'un postdoc (à recruter en 2018). Bien sûr, les recherches à mener dans ce contexte seront aussi utiles dans l'expertise que je pourrais mettre au service d'organismes tels l'ISO ou l'OGC qui s'intéressent actuellement à l'interopérabilité BIM/SIG.

Le projet de recherche décrit à la section 3.4 se base sur mes travaux, présents et passés, de même que sur mon expérience acquise à travers mes activités et projets. Il s'agit de continuer à travailler avec les technologies sémantiques pour améliorer comment les ordinateurs interprètent les données qu'ils manipulent, certes avec une nouvelle approche (définition de granularités dans les connaissances), mais toujours en essayant de trouver le meilleur compromis entre l'expressivité des modèles sémantiques utilisés et le niveau d'efficacité de leur implémentation. L'idée à long terme est qu'en spécifiant des granules de connaissance, de manière suffisamment haut niveau (avec des méta-méta-modèles par exemple), pourrait permettre de ***passer de données liées à des connaissances liées***. Un haut niveau de spécification de ces granules pourrait aider dans leur découverte, composition voire même leur intégration par des agents informatiques et ce sans scénarios établis au prélable.

Enfin, un nouveau domaine d'application a émergé suite au dépôt du projet H2020 BigSTEP et suite au démarrage du projet ANR McBIM : il s'agit de l'annotation sémantique de flux de données capteurs et de leur intégration dans un SI (qu'il soit géographique ou pas) afin de pouvoir raisonner sur ces données et notamment de pouvoir en extraire des connaissances. En effet, avec la standardisation par le W3C de l'ontologie SSN (Semantic Sensor Networks), au travers de la collaboration initiée en 2017 avec Maxime Lefrançois (Mines Saint-Etienne) et grâce à la collaboration avec le LAAS dans le projet ANR McBIM, je souhaite prendre en compte dans mes recherches les contraintes et particularités du domaine des capteurs.

A terme, je vise d'inscrire mes recherches dans le domaine des systèmes basés sur l'Intelligence Artificielle (AI-based systems). Leur champ d'application s'étend aujourd'hui de la recommandation de films à de l'assistance à la conduite en passant par la sécurité aux frontières. Le vaste panorama d'applications pose de nouveaux défis scientifiques, notamment dans l'expression et la quantification du niveau de confiance qu'a l'utilisateur (e.g. par rapport à leur performance, leur intégrité, les déductions faites, leurs capacités de corriger des erreurs ou encore de les prévoir).

4 Bibliographie

4.1 Mes publications

4.1.1 Revues Internationales à Comité de Lecture

[RIS1] Tarcisio Mendes De Farias, **Ana Roxin**, Christophe Nicolle. A rule-based methodology to extract building model views. *Automation in Construction*, Elsevier, 2018, pp.214-229. 10.1016/j.autcon.2018.03.035
SCImago journal rank : Q1

[RIS2] Muhammad Arslan, Christophe Cruz, **Ana Roxin**, Dominique Ginhac, Spatio-temporal Analysis of Trajectories for Safer Construction Sites. *Journal of Smart and Sustainable Built Environment (SASBE)* (CIB Encouraged Journal), Emerald, 2018. SCImago journal rank : Q3, Scopus CieScore: 0.63, Scopus CiteScoreTracker 2017: 1.00

[RIS3] Pieter Pauwels, Tarcisio Mendes De Farias, Chi Zhang, **Ana Roxin**, Jakob Beetz, et al. A performance benchmark over semantic rule checking approaches in construction industry. *Advanced Engineering Informatics*, Elsevier, 2017, 33, pp.68-88. 10.1016/j.aei.2017.05.001. SCImago journal rank : Q1

[RIS4] Timo Homburg, Claire Prudhomme, Frank Boochs, **Ana Roxin**, Christophe Cruz. Integration, quality assurance and usage of geospatial data with semantic tools. *gis.Science - Die Zeitschrift für Geoinformatik*, Wichmann, Septembre 2017. hal-01684684. SCImago journal rank : Q4

[RIS5] Tarcisio Mendes De Farias, **Ana Roxin**, Christophe Nicolle. SWRL rule-selection methodology for ontology interoperability. *Data & Knowledge Engineering*, 105, pp.53-72. 10.1016/j.datak.2015.09.001
Journal impact factor in 2015: 1.12. Acceptance rate: 7 %. SCImago journal rank : Q1

[RIS6] Anett Hoppe, **Ana Roxin**, Christophe Nicolle. Semantic User Profiling for Digital Advertising, in *Journal of Economic Computation and Economic Cybernetics Studies and Research*, Mar 2015, 49 (3/2015), pp.267-286. hal-01202762 - Impact factor in 2014/2015: 0.406. SCImago journal rank: Q3

[RIS7] Anett Hoppe, **Ana Roxin**, Christophe Nicolle. Ontology-based Integration of Web Navigation for Dynamic User Profiling. 15 pages, in *Journal "Informatica Economică"*, Special issue on Business Ontologies, Vol 19 (01/2015) hal-01165191 Open Access Journal ISSN 1453-1305 / EISSN 1842-8088

[RIS8] Julien Henriet, Pierre-Emmanuel Leni, Remy Laurent, **Ana Roxin**, Brigitte Chebel-Morello, et al. Adapting numerical representations of lung contours using Case-Based Reasoning and Artificial Neural Networks. In: Agudo B.D., Watson I. (eds) *Case-Based Reasoning Research and Development. ICCBR 2012. Lecture Notes in Computer Science*, vol 7466. Springer, Berlin, Heidelberg, 2013, pp.137-151.
10.1007/978-3-642-32986-9_12

4.1.2 Chapitres de livres

[CH1] Nicolas Bus, Muhammad Fahad, Bruno Fies, **Ana Roxin**, Semantic topological querying for compliance checking. e*Work and eBusiness in Architecture, Engineering and Construction: ECPPM 2018*, CRC Press, August 2018, ISBN 9781138584136.

[CH2] Claire Prudhomme, **Ana Roxin**, Christophe Cruz, Frank Boochs. Towards the Design of Respond Action in Disaster Management Using Knowledge Modeling. In: Dokas I., Bellamine-Ben Saoud N., Dugdale J., Díaz P. (eds) *Information Systems for Crisis Response and Management in Mediterranean Countries*. ISCRAM-med 2017. Lecture Notes in Business Information Processing, vol 301. Springer, Cham

[CH3] Timo Homburg, Claire Prudhomme, Falk Würriehausen, Ashish Karmacharya, Frank Boochs, **Ana Roxin**, Christophe Cruz. Interpreting Heterogeneous Geospatial Data Using Semantic Web Technologies. In: Gervasi O. et al. (eds) *Computational Science and Its Applications - ICCSA 2016*.

Lecture Notes in Computer Science, vol 9788. Springer, Cham pp. 240-255, 2016, ISBN 978-3-319-42110-0. DOI:10.1007/978-3-319-42111-7_19

[CH4] Pieter Pauwels, **Ana Roxin**. SimpleBIM: From full ifcOWL graphs to simplified building graphs. e*Work and eBusiness in Architecture, Engineering and Construction: ECPPM 2016*, CRC Press, pp.11-18, 2016, ISBN 9781138032804. hal-01389818

[CH5] Udo Kannengiesser, **Ana Roxin**. An agile process modelling approach for BIM projects. e*Work and eBusiness in Architecture, Engineering and Construction: ECPPM 2016*, CRC Press, pp.99-107, 2016, ISBN 9781138032804. hal-01389816

[CH6] Anett Hoppe, **Ana Roxin**, Christophe Nicolle. Customizing Semantic Profiling for Digital Advertising. *On the Move to Meaningful Internet Systems: OTM 2014 Workshops*, 8842 (0302-9743), Springer Berlin Heidelberg, pp 469-478, 2014, Lecture Notes in Computer Science, 978-3-662-45549-4. DOI : 10.1007/978-3-662-45550-0_47.

[CH7] Tarcisio Mendes de Farias, **Ana Roxin**, Thomas Durif, Florian Orpelière, Christophe Nicolle. Enrichissement sémantique d'un fichier IFC pour une extraction partielle dynamique. Édité par Jean-Claude Bignon, Gilles Halin, Sylvain Kubicki. *Interaction(s) des maquettes numériques*, Presses universitaires de Nancy, 334 p., 2014, ISBN : 9782814301719. hal-01072014

[CH8] Ioan Roxin, **Ana Roxin**. Outils sémantiques pour diminuer l'entropie dans les réseaux socionumériques. *IDENTITÉS NUMÉRIQUES*, Groupe L'Harmattan, pp.94-112, 2013, 978-2-343-01999-4. hal-00945144

[CH9] **Ana Roxin**. L'information Géographique. Dans *Géopositionnement et Mobilités*, édité par: UTBM. Publié en Août 2009, ISBN: 978-2-014279-40-6.

[CH10] **Ana Roxin**. Communications inter-véhicules. Dans *Géopositionnement et Mobilités*, édité par: UTBM. Publié en Août 2009, ISBN: 978-2-014279-40-6.

4.1.3 Conférences internationales avec comité de lecture

[CI1] Nicolas Bus, **Ana Roxin**, Guillaume Picinbono, Muhammad Fahad, Towards French smart building code : Compliance checking based on semantic rules. *LDAC2018 – 6th Linked Data in Architecture and Construction Workshop (LDAC2018)*, Jun 2018, London, United-Kingdom. Disponible en ligne : http://ceur-ws.org/Vol-2159/01paper.pdf

[CI2] Timo Homburg, Frank Boochs, Christophe Cruz, **Ana Roxin**. Map Change Prediction for Quality Assurance. In : *Adjunct Proceedings of the 14th International Conference on Location Based Services*. ETH Zurich, 2018. p. 194-200. Disponible en ligne : https://www.research-collection.ethz.ch/handle/20.500.11850/225617

[CI3] Muhammad Arslan, **Ana Roxin**, Christophe Cruz, Dominique Ginhac. A Review on Applications of Big Data for Disaster Management. *13th IEEE International Conference on Signal Image Technology & Internet-based Systems*. IEEE SITIS 2017, Jaipur, India.

[CI4] Muhammad Arslan, Christophe Cruz, **Ana Roxin**, Dominique Ginhac. Using Spatio-temporal Trajectories to Monitor Construction Sites for Safety Management. *9th International Conference on Information Management and Engineering (ICIME 2017),* Oct 2017, Barcelone, Spain. ACM Press, 10.1145/3149572.3149600

[CI5] Claire Prudhomme, **Ana Roxin**, Christophe Cruz, Frank Boochs. Towards the Design of Respond Action in Disaster Management Using Knowledge Modeling. *Information Systems for Crisis Response and Management in Mediterranean Countries*, Oct 2017, Xanthi, Greece. hal-01684617

[CI6] Claire Prudhomme, Timo Homburg, Jean-Jacques Ponciano, Frank Boochs, **Ana Roxin**, et al. Automatic Integration of Spatial Data into the Semantic Web. *WebIST 2017*, Apr 2017, Porto, Portugal. hal-01493390

[CI7] Pieter Pauwels, Tarcisio Mendes de Farias, Zhang Chi, **Ana Roxin**, Beetz Jakob, et al. Querying and reasoning over large scale building data sets: an outline of a performance benchmark. *The International Workshop on Semantic Big Data (SBD '16) ACM in conjunction with ACM SIGMOD 2016*, Jul 2016, San Francisco United States. 2016. 10.1145/2928294.2928303. hal-01329400

[CI8] Tarcisio Mendes de Farias, **Ana Roxin**, Christophe Nicolle. A Federated Approach for Interoperating AEC/FM Ontologies. *LDAC2016 - 4th Linked Data in Architecture and Construction Workshop*, Jun 2016, Madrid, Spain. hal-01329726

[CI9] Claire Prudhomme, **Ana Roxin**, Christophe Cruz, Frank Boochs. Semantic Catalogue to Manage Data Sources in Disaster Management System. *15th International Conference on Informatics in Economy (IE 2016)*, pp.225-230, Jun 2016, Cluj Napoca, Romania. hal-01438584

[CI10] Tarcisio Mendes de Farias, **Ana Roxin**, Christophe Nicolle. COBieOWL, an OWL ontology based on COBie standard. In proceedings of *On the Move to Meaningful Internet Systems*: OTM 2015 Conferences, Confederated International Conferences: CoopIS, ODBASE, and C&TC 2015, Oct 2015, Rhodes, Greece. Lecture Notes in Computer Science 9415, Springer 2015. hal-01202742 Acceptance rate: 23%

[CI11] Tarcisio Mendes de Farias, **Ana Roxin**, Christophe Nicolle. IfcWoD, Semantically Adapting IFC Model Relations into OWL Properties. *In proceedings of the 32nd CIB W78 Conference on Information Technology in Construction*, Oct 2015, Eindhoven, Netherlands. hal-01174524 Acceptance rate (including abstract phase): 45%

[CI12] Tarcisio Mendes de Farias, **Ana Roxin**, Christophe Nicolle. FOWLA, A Federated Architecture for Ontologies. *International Web Rule Symposium (RuleML 2015)*, Aug 2015, Berlin, Germany. LNCS 9202, pp.97-111, 2015, In N. Bassiliades et al. (Eds.): Rule Technologies: Foundations, Tools, and Applications. hal-01174510 Main conference's full paper acceptance rate: 40%

[CI13] Tarcisio Mendes de Farias, **Ana Roxin**, Christophe Nicolle. Semantically Adapting IFC Model Relations into OWL Properties. *3rd International Workshop on Linked Data in Architecture and Construction (LDAC)*, Jul 2015, Eindhoven, Pays-Bas. Workshop Report. hal-01202808

[CI14] Anett Hoppe, **Ana Roxin**, Christophe Nicolle. Automatic User Profile Mapping To Marketing Segments In A Big Data Context. *14th International Conference on Informatics in Economy (IE 2015)*, Apr 2015, Bucharest, Romania. hal-01165211

[CI15] Tarcisio Mendes de Farias, **Ana Roxin**, Christophe Nicolle. Semantic Web Technologies for Implementing Cost Effective and Interoperable Building Information Modeling. *The 14th International Conference on Informatics in Economy (IE 2015)*, Apr 2015, Bucharest, Romania. hal-01174498

[CI16] Anett Hoppe, **Ana Roxin**, Christophe Nicolle. Customizing Semantic Profiling for Digital Advertising. *On the Move to Meaningful Internet Systems: OTM 2014 Workshops*, October 2014, Amantea, Italy. Lecture Notes in Computer Science 8842 (0302-9743), Springer Berlin Heidelberg, pp 469-478, ISBN: 978-3-662-45549-4. Acceptance rate: 40% hal-01094420

[CI17] Tarcisio Mendes de Farias, **Ana Roxin**, Christophe Nicolle. A Rule Based System for Semantical Enrichment of Building Information Exchange. Theodore Patkos, Adam Wyner and Adrian Giurca. *RuleML 2014*, Aug 2014, Prague, Czech Republic. Vol-1211, pp.8, 2014. hal-01068809

[CI18] Tarcisio Mendes de Farias, **Ana Roxin**, Christophe Nicolle, Thomas Durif, Florian Orpelière. Enrichissement sémantique d'un IFC pour une extraction partielle dynamique. *Séminaire de Conception Architecturale Numérique*, Juin 2014, Luxembourg, Luxembourg. Actes du 6ème Séminaire de Conception Architecturale Numérique. hal-00978429 Full paper acceptance rate: 60%

[CI19] Anett Hoppe, **Ana Roxin**, Christophe Nicolle. Dynamic Semantic User Profiling from Implicit Web Navigation Data. *13th International Conference on Informatics in Economy*, May 2014, Bucarest, Romania. ISSN: 2247 - 1480, 2014. hal-00978436

[CI20] Anett Hoppe, **Ana Roxin**, Christophe Nicolle. Dynamic, Behavior-Based User Profiling Using Semantic Web Technologies in a Big Data Context. Demey, YanTang and Panetto, Hervé. *On the Move to Meaningful Internet Systems: OTM 2013 Workshops*, Sep 2013, Graz, Austria. Lecture Notes in Computer Science 8186. Acceptance rate: 30% hal-00850608

[CI21] Julien Henriet, Pierre-Emmanuel Leni, Remy Laurent, **Ana Roxin**, Brigitte Chebel-Morello, et al. Adapting numerical representations of lung contours using Case-Based Reasoning and Artificial Neural Networks. Proceedings of the *20th International Conference on Case-Based Reasoning (ICCBR)*, LNCS 7466, B. Díaz Agudo and I. Watson (Eds.), Springer, Heidelberg, 2012, pp.137-151. hal-00714584

[CI22] Ioan Szilagyi, Radu Balog-Crisan, **Ana Roxin**, Ioan Roxin. Ontologies and Knowledge Aggregation in the Active Semantic Learning System. *11th IEEE International Conference on Advanced Learning Technologies (ICALT 2011)*, 6-8 July 2011, Athens, USA.

[CI23] Ahmed Nait-Sidi-Moh, Maxime Wack, Jafaar Gaber, **Ana Roxin**. Modeling and implementation of a pervasive e-learning application. *International Conference on Multimedia Computing and Systems (ICMCS)*, 7-9 April 2011, Ouarzazate, Morocco.

[CI24] Ioan Szilagyi, Radu Balog-Crisan, Ioan Roxin, **Ana Roxin**. Kernel for a Semantic Learning Platform with Adapted Suggestions. *10th IEEE International Conference on Advanced Learning Technologies (ICLAT'10)*, p.400-402, 5-7 July 2010, Sousse, Tunisia.

[CI25] **Ana Roxin**, Christophe Dmuez, Nathanael Cottin, et al. TransportML: A middleware for location-based services collaboration. IEEE 3rd International Conference on *New Technologies, Mobility and Security (NTMS)*, p. 1-6. 2009.

4.1.4 Conférences nationales avec comité de lecture

[CN1] **Ana Roxin**. Plateforme d'apprentissage à distance en radioprotection médicale. *28e Journées des Laboratoires Associés en Radioprotection et Dosimétrie (L.A.R.D.)*, 10-12 Octobre 2011, Toulouse.

[CN2] **Ana Roxin**, Régine Gschwind, Libor Makovicka, Ioan Roxin, Julien Henriet, Eric Martin. CEFOR@D - Plateforme de formation à distance pour la radioprotection médicale. *8e Congrès National de Radioprotection SFRP (Société Française de Radioprotection)*, 21-23 Juillet 2011, Tours. http://www.sfrp.asso.fr/IMG/pdf/S4a-Roxin.pdf

4.2 Bibliographie

(W3C), W.W.W.C., 2013. RIF Framework for Logic Dialects (Second Edition).

(w3c), W.W.W.C., 2015. Semantic web : Inference. Available at: http://www.w3.org/standards/semanticweb/inference.

Ackoff, R., 1989. From data to wisdom. *Journal of Applied Systems Analysis, vol*, 16, pp.3–9.

Adar, E., Karger, D. & Stein, L.A., 1999. Haystack : Per-user information environments. In *Proceedings of the eighth international conference on Information and knowledge management*. ACM, pp. 413–422.

Agichtein, E., Brill, E. & Dumais, S., 2006. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 19–26.

Agostinho, C. et al., 2015. Towards a sustainable interoperability in networked enterprise information systems: Trends of knowledge and model-driven technology. *Computers in Industry*.

Agostinho, C. et al., 2011. Tuple-based semantic and structural mapping for a sustainable interoperability. In *Technological Innovation for Sustainability*. Springer, pp. 45–56.

Aitsi-selmi, A., Egawa, S. & Sasaki, H., 2015. The Sendai framework for disaster risk reduction: Renewing the global commitment to people's resilience, health, and well-being. *International Journal of Disaster Risk Science*, 6(2), pp.164–176.

Allemang, D. & Hendler, J., 2011. Semantic web for the working ontologist: effective modeling in RDFS and OWL.

Alliance, O.D., 2013. *Open Design Specification for .dwg files*, Available at: http://opendesign.com/files/guestdownloads/OpenDesign_Specification_for_.dwg_files.pdf.

Angele, J., 2014. OntoBroker. *Semantic Web*, 5(3), pp.221–235.

Apt, K.R., 1997. From logic programming to Prolog.

Arapakis, I. et al., 2009. Integrating facial expressions into user profiling for the improvement of a multimodal recommender system. In *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, pp. 1440–1443.

Arthur O'Sullivan, S.P. Steven Sheffrin, 2016. *Macroeconomics: Principles, Applications, and Tools* 9th ed. P. Education, ed., New Jersey: Pearson Education.

Autodesk, I., 2010. DXF Reference.

Baader, F. et al., 2003. *The description logic handbook : theory, implementation, and applications*, Cambridge University Press.

Bac, M. et al., 2012. A-trader - consulting agent platform for stock exchange gamblers. In *FedCSIS*. pp. 963–968.

Bak, J., Nowak, M. & Jedrzejek, C., 2014. RuQAR: Reasoning Framework for OWL 2 RL Ontologies. In E. Blomqvist et al., eds. *The Semantic Web: ESWC 2014 Satellite Events*. Springer, Cham.

Barbau, R. et al., 2012. OntoSTEP: Enriching product model data using ontologies. C*omputer-Aided Design*, 44(6), pp.575–590.

Barg, M. et al., 2000. Problem-based learning for foundation computer science courses. C*omputer Science Education*, 10(2), pp.109–128.

Baroglio, C. et al., 2008. *Reasoning Web: 4th International Summer School 2008, Venice Italy, September 7-11, 2008, Tutorial Lectures*, Springer.

Barton, D. & Court, D., 2012. Making advanced analytics work for you. *Harvard Business Review*, 90(10), pp.78–83.

Bazoun, H. et al., 2013. Transformation of Extended Actigram Star to BPMN2. 0 in the frame of Model Driven Service Engineering Architecture. In *Proceedings of the Symposium on theory of Modeling and Simulation*.

Bechhofer, S., 2004. Hoolet.

Beetz, J., Van Leeuwen, J. & De Vries, B., 2009. IfcOWL: A case of transforming EXPRESS schemas into ontologies. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 23(01), pp.89–101.

Bellifemine, F., Poggi, A. & Rimassa, G., 1999. *JADE - A FIPA-compliant agent framework*, CSELT, Tech. Rep.

Benslimane, D. et al., 2006. Contextual Ontologies. In T. Yakhno & E. J. Neuhold, eds. *Advances in Information Systems: 4th International Conference, ADVIS 2006, Izmir, Turkey, October 18-20,2006. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 168–176.

Berners-Lee, T., 2006. Linked Data. Available at: http://www.w3.org/DesignIssues/LinkedData.html.

Berners-lee, T. et al., 2008. N3Logic: A Logical Framework for the World Wide Web. *Theory Pract. Log. Program.*, 8(3), pp.249–269. Available at: http://dx.doi.org/10.1017/S1471068407003213.

Bittner, T. & Smith, B., 2001. A unified theory of granularity, vagueness and approximation.

Bittner, T. & Smith, B., 2003. Granular spatio-temporal ontologies. In *Proceedings of the AAAI Spring Symposium on Foundations and Applications of Spatio-Temporal Reasoning (FASTR)*. Menlo Park, CA: AAAI Press, pp. 12–17.

Bouchachia, A., Lena, A. & Vanaret, C., 2014. Online and interactive self-adaptive learning of user profile using incremental evolutionary algorithms. *Evolving Systems*, 5(3), pp.143–157.

Bouquet, F. et al., 2015. Formalismes de description des modèles agent. In *Simulation spatiale à base d'agents avec NetLogo 1 : introduction et bases*. pp. 37–72.

Bowers, S., Madin, J.S. & Schildhauer, M.P., 2010. Owlifier: Creating OWL-DL ontologies from simple spreadsheet-based knowledge descriptions. *Ecological Informatics*, 5(1), pp.19–25.

Brooks, R.A., 1991. Intelligence without Reason. In *in Proceedings of the 1991 International Joint Conference on Artificial Intelligence*. pp. 569–595.

Bruce, K., 2011. Lecture 18: Axiomatic Semantics & Type safety, CSCI 131, Pomona College.

Brynjolfsson, E. & McAfee, A., 2012. *Race against the machine: how the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy* D. F. Press, ed., Digital Frontier Press.

BuildingSMART, 2016. MVD Overview summary.

Bürger, T. & Simperl, E.P.B., 2008. Measuring the benefits of ontologies. In Z. Tari, ed. *On the Move to Meaningful Internet Systems: OTM 2008 Workshops*. Springer, pp. 584–594.

Butenkov, S. et al., 2017. Granular computing models and methods based on the spatial granulation. *Procedia Computer Science*, 103, pp.295–302.

Calegari, S. & Ciucci, D., 2010. Granular computing applied to ontologies. *International Journal of Approximate Reasoning*, 51(4), pp.391–409. Available at: https://doi.org/10.1016/j.ijar.2009.11.006.

Carpenter, G.A. & Grossberg, S., 1987. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, 37(1), pp.54–115.

Carreira, R. et al., 2004. Evaluating adaptive user profiles for news classification. In *Proceedings of the 9th international conference on Intelligent user interfaces*. ACM, pp. 206–212.

Castano, S., Ferrara, A. & Montanelli, S., 2006. Ontology-based interoperability services for semantic collaboration in open networked systems. *Interoperability of Enterprise Software and Applications*, pp.135–146.

Castellano, G., Fanelli, A. & Torsello, M., 2007. Web user profiling using relational fuzzy clustering. *Information Sciences*, pp.1433–1439.

Challam, V., Gauch, S. & Chandramouli, A., 2007. Contextual search using ontology-based user profiles. In *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, pp. 612–617.

Charalabidis, Y. et al., 2010. A review of electronic government interoperability frameworks: patterns and challenges. *International Journal of Electronic Governance*, 3(2), pp.189–221.

Chen, D., Doumeingts, G. & Vernadat, F., 2008. Architectures for enterprise integration and interoperability: Past, present and future. *Computers in industry*, 59(7), pp.647–659.

Chen, D. & Vernadat, F., 2004. Standards on enterprise integration and engineering—state of the art. *International Journal of Computer Integrated Manufacturing*, 17(3), pp.235–253.

Chen, D. & Vernadat, F.B., 2003. Enterprise interoperability: A standardisation view. In *Enterprise Inter- and Intra-Organizational Integration*. Springer, pp. 273–282.

Chen, Y.-J., Chen, Y.-M. & Chu, H.-C., 2009. Development of a mechanism for ontology-based product lifecycle knowledge integration. *Expert Systems with Applications*, 36(2, Part 2), pp.2759–2779.

Chirita, P. -a., Firan, C.S. & Nejdl, W., 2007. Personalized query expansion for the web. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 7–14.

Chu, W. & Park, S. -t., 2009. Personalized recommendation on dynamic content using predictive bilinear models. In *Proceedings of the 18th international conference on World wide web*. pACM, pp. 691–700.

Claypool, M. et al., 2001. Implicit interest indicators. In *Proceedings of the 6th international conference on Intelligent user interfaces*. ACM, pp. 33–40.

Confalonieri, R., Iñan, H. & Palau, M., 2012. Handling uncertain user preferences in a context-aware system. In *Advances on Computational Intelligence - 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Catania, Italy: IPMU 2012, pp. 88–97.

Correndo, G. et al., 2010. SPARQL query rewriting for implementing data integration over linked data. In *Proceedings of the 2010 EDBT/ICDT Workshops*. ACM, p. 4.

Corry, E. et al., 2015. A performance assessment ontology for the environmental and energy management of buildings. *Automation in Construction*, *57*, pp.249–259.

Dean, M. et al., 2006. OWL web ontology language reference. W3C Recommendation, 2004.

Dibley, M.J., 2011. An intelligent system for facility management.

Domingue, J., Fensel, D. & Hendler, editors J. A., 2011. *Handbook of Semantic Web Technologies*, Springer.

Ducq, Y., Chen, D. & Alix, T., 2012. Principles of servitization and definition of an architecture for model driven service system engineering. In *Enterprise Interoperability*. Springer, pp. 117–128.

Dudev, M. et al., 2008. Personalizing the search for knowledge. In *PersDB 2nd International Workshop on Personalized Access, Profile Management, and Context Awareness*. Auckland, New Zealand, pp. 1–8.

East, B., 2014. Construction-operations building information exchange (cobie). Available at: http://www.wbdg.org/resources/cobie.php.

Eastman, C. et al., 2011. *BIM Handbook: A Guide to Building Information Modeling for Owners, Managers Designers, Engineers and Contractors (2nd ed.)*, Wiley.

Eastman, C.M. et al., 2008. *A Guide to Building Information Modeling for Owners, Managers, Architects, Engineers, Contractors, and Fabricators*, Hoboken, NJ: John Wiley and Sons.

Economist, T., 2017a. Fuel of the future. *The Economist*.

Economist, T., 2017b. The world's most valuable resource. *The Economist*.

Efron, B., 1979. Bootstrap methods : another look at the jackknife. *The annals of Statistics*, pp.1–26.

Elenius, D., 2012. SWRL-IQ: A Prolog-based Query Tool for OWL and SWRL. In *OWLED*.

Ericksons S., H. Rothberg, 2014. Big Data and Knowledge Management: Establishing a Conceptual Foundation. *The Electronic Journal of Knowledge Management*, 12(2), pp.108–116.

Euzenat, J., 2011. Networks of ontologies and alignments.

Euzenat, J. & Shvaiko, P., 2013. *Ontology matching* Second Edition, Springer-Verlag Berlin Heidelberg, Germany.,

eXelate, 2012. *Employing adaptive audience intelligence*, Available at: http://exelate.com/white-papers/employing-adaptive-audienceintelligence-2/.

Eyharabide, V. & Amandi, A., 2012. Ontology-based user profile learning. *Appl. Intell.*, 36(4), pp.857–869.

Ferguson, I.A., 1992. Touring machines: Autonomous agents with attitudes. *Computer*, 25(5), pp.51–55.

Forum, W.E., 2015. *Deep Shift – Technology Tipping Points and Societal Impacts (Geneva: World Economic Forum*, Global Agenda Council on the Future of Software & Society. Available at: http://www3.weforum.org/docs/WEF_GAC15_Technological_Tipping_Points_report_2015.pdf.

Frey, C.B. & Osborne, M.A., 2013. The future of employment: how susceptible are jobs to computerisation? Available at: https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf.

Gabbay, D.M. et al., 2003. Modal logic basics. *Studies in Logic and the Foundations of Mathematics*, 148, pp.3–40.

Gallier, J.H., 2015. *Logic for computer science: foundations of automatic theorem proving*, Courier Dover Publications.

Gao, G. et al., 2015. A query expansion method for retrieving online BIM resources based on Industry Foundation Classes. *Automation in Construction*, 56, pp.14–25.

Gasparetti, F. & Micarelli, A., 2005. User profile generation based on a memory retrieval theory. In *Proceedings of the 1st International Workshop on Web Personalization, Recommender Systems and Intelligent User Interfaces (WPRSIUI'05)*. Citeseer, pp. 59–68.

Gauch, S. et al., 2007. User profiles for personalized information access. In *The adaptive web*. Springer, pp. 54–89.

Gehre, A. et al., 2006. EU FP6 Inteligrid: Interoperability of virtual organizations on a complex semantic grid. dresden : Available at: https://cordis.europa.eu/publication/rcn/9411_en.html.

Géry, M. & Haddad, H., 2003. Evaluation of web usage mining approaches for user's next request prediction. In *Proceedings of the 5th ACM international workshop on Web information and data management*. ACM, pp. 74–81.

Ghorab, M.R. et al., 2013. Personalised information retrieval : survey and classification. *User Modeling and User-Adapted Interaction*, 23(4), pp.381–443.

Girard, B. & Khamassi, M., 2016. *Coopération d'algorithmes d'apprentissage par renforcement multiples, Techniques de l'ingénieur Réf : S7793 v1, 10 déc. 2016*, Techniques de l'ingénieur. Available at: https://www.techniques-ingenieur.fr/base-documentaire/automatique-robotique-th16/perception-planification-et-interface-en-robotique-42622210/cooperation-d-algorithmes-d-apprentissage-par-renforcement-multiples-s7793/.

Glimm, B. et al., 2014. HermiT: an OWL 2 reasoner. *Journal of Automated Reasoning*, 53(3), pp.245–269.

Golemati, M. et al., 2007. Creating an ontology for the user profile : Method and applications. In *Proceedings of the First RCIS Conference*. pp. 407–412.

Golovchinsky, G., Price, M.N. & Schilit, B.N., 1999. From reading to retrieval : freeform ink annotations as queries. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 19–25.

Goodman, N. & Leonard, H., 1940. The calculus of individuals and its uses ». *The Journal of Symbolic Logic, Vol*, 5(2), pp.45–55.

Grcar, M., Mladenic, D. & Grobelnik, M., 2005. User profiling for interest-focused browsing history. In *In SIKDD 2005 at Multiconference IS 2005*.

Grenon, P. & Smith, B., 2004. SNAP and SPAN: Prolegomenon to geodynamic ontology. *Spatial Cognition and Computation*, 1, pp.69–105.

Grimm, S. et al., 2011. Ontologies and the semantic web. *Handbook of Semantic Web Technologies*, pp.507–579.

Gupta, P. et al., 2013. Wtf : The who to follow service at twitter. In *Proceedings of the 22Nd International Conference on World Wide Web*. Republic and Canton of Geneva, Switzerland: WWW '13, pp. 505–514.

Haarslev, V. et al., 2012. The RacerPro knowledge representation and reasoning system. *Semantic Web*, 3(3), pp.267–277. Available at: http://www.semantic-web-journal.net/sites/default/files/swj109_3.pdf.

Haarslev, V., Pai, H.-I. & Shiri, N., 2008. Uncertainty reasoning for ontologies with General TBoxes in description logic. *Uncertainty Reasoning for the Semantic Web I*, pp.385–402.

Halevy, A., Norvig, P. & Pereira, F., 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2), pp.8–12.

Halevy, A.Y. et al., 2005. Enterprise information integration: successes, challenges and controversies. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. ACM, pp. 778–787.

Harris, S. & Seaborne, A., 2013. SPARQL 1.1 query language for RDF. Available at: https://www.w3.org/TR/sparql11-query/.

Hawe, G.I. et al., 2012. Agent-based simulation for large-scale emergency response: A survey of usage and implementation. *ACM Computing Surveys (CSUR)*, 45(1).

Hayes, P.J. & Patel-Schneider, P.F., 2014. RDF 1.1 Semantics - W3C Recommendation 25 February 2014. Available at: https://www.w3.org/TR/rdf11-mt/.

Von Hessling, A., Kleemann, T. & Sinner, A., 2005. Semantic user profiles and their applications in a mobile environment. *Fachberichte Informatik*, 1(2).

Hoang, N.V. & Törma, S., 2014. *Opening BIM to the Web IFC-to-RDF Conversion Software*, Available at: http://rym.fi/results/opening-bim-to-the-web-ifcto-rdf-conversion-software/.

Hobbs, J.R., 1985. Granularity. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence*. Los Angeles, CA.

Hoppe, A., 2016. *Enhancing Ontology-Based User Profiling by Uncertainty Handling - An Application to the Digital Advertising Domain*. Université de Bourgogne (uB).

Horrocks, I. et al., 2004. SWRL : A semantic web rule language combining OWL and RuleML - W3C Member Submission 21 May 2004. Available at: https://www.w3.org/Submission/SWRL/.

Horvitz, E. et al., 1998. The lumiere project : Bayesian user modeling for inferring the goals and needs of software users. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc, pp. 256–265.

IBM, 2014. *Understanding Big Data so you can act with confidence*, IBM Corporation. Available at: https://tdwi.org/ /media/F2AF8C2A06824E38A9A908770AF2BD9C.PDF.

Inc., C., 2016. The stardog manual.

ISO/IEC, J.S.W., 1991. SUMM Semantic Unification Meta Model.

ISO10027, 1990. ISO 10027:1990 - Information technology – Information Resource Dictionary System (IRDS) framework.

ISO10303-21, 2002. Product data representation and exchange - Part 21 (ISO 10303-21:2002).

ISO14258, 1999. Industrial automation systems and integration — Concepts and rules for enterprise models.

ISO16739, 2013. (ISO 16739:2013) Classes de fondation d'industrie (IFC) pour le partage des données dans le secteur de la construction et de la gestion des installations.

ISO17027, 2014. *Évaluation de la conformité — Vocabulaire lié à la compétence des personnes utilisée pour la certification de personnes*,

ISO17261, 2012. Systèmes intelligents de transport – Identification automatique des véhicules et des équipements – Architecture et terminologie du transport intermodal des marchandises.

ISO29481-1, 2016. (ISO 29481-1:2016) - Building information models - Information delivery manual - Part 1: Methodology and format.

ISO29481-3, 2010. (ISO 29481-3:2010) - Building information models - Information delivery manual - Part 3: Model View Definition.

ISO772, 2011. Hydrométrie – Vocabulaire et symboles.

ISO8000-110, 2009. (ISO 8000-110:2009) Data quality – Part 110: Master data: Exchange of characteristic data: Syntax, semantic encoding, and conformance to data specification.

Jakubczyc, J.A. & Owoc, M.L., 2011. Contextual knowledge granularity. In *Proceedings of Informing Science & IT Education Conference (InSITE)*. pp. 259–268.

Jang, M. & Sohn, J.-C., 2004. Bossam: An extended rule engine for OWL inferencing. *Rules and Rule Markup Languages for the Semantic Web*, pp.128–138.

Jena, A., 2016. *Reasoners and rule engines : Jena inference support*, Available at: https://jena.apache.org/documentation/inference/.

Jupp, S. et al., 2010. Populous: A Tool for Populating Templates for OWL Ontologies. *SWAT4LS*.

Kaplanski, P. & Weichbroth, P., 2017. Cognitum Ontorion: Knowledge Representation and Reasoning System. In P.-P. T. & O. C. Mach-Król M., eds. *New Ideas from Ongoing Research. Studies in Computational Intelligence, vol 658*. Advances in Business ICT: Springer, Cham.

Keet, C.M., 2008. *A Formal Theory of Granularity - Towards enhancing biological and applied life sciencesinformation systems with granularity*. Faculty of Computer Science, Free University of Bozen - Bolzano.

Kelly, D. & Teevan, J., 2003. Implicit feedback for inferring user preference : a bibliography. *ACM SIGIR Forum*, 37, pp.18–28.

Kemp, J.L.C., 1999. *Fractal organising of knowledge intensive organisations*, Erasmus Universiteit.

Kifer, M., 2005. Rules and ontologies in f-logic. In *Reasoning Web*. Springer, pp. 22–34.

Kim, Y.S. et al., 2005. Development of a recommender system based on navigational and behavioral patterns of customers in e-commerce sites. *Expert Systems with Applications*, 28(2), pp.381–393.

Kleppe, A.G., Warmer, J.B. & Bast, W., 2003. *MDA explained: the model driven architecture: practice and promise*, Addison-Wesley Professional.

Knublauch, H., 2011. SPIN - Modeling Vocabulary.

Korczak, J., Hernes, M. & Bac, M., 2013. Risk avoiding strategy in multiagent trading system. *FedCSIS*, pp.1119–1126.

Krima, S. et al., 2009. Ontostep: Owl-dl ontology for step. In *Proceedings of 6 th International Conference on Product Lifecycle Management*. Citeseer.

Kripke, S., 1980. *Naming and necessity*, Harvard University Press.

Krötzsch, M., Simancik, F. & Horrocks, I., 2012. A description logic primer. Co*RR*, abs/1201.4089.

Kurzweil, R., 2005. *The Singularity is Near*, Penguin Books.

Kwon, O. & Kim, J., 2009. Concept lattices for visualizing and generating user profiles for context-aware service recommendations. *Expert Systems with Applications*, 36(2 - Part 1), pp.1893–1902.

LDWG, B., 2015. IFC2X3_TC1 - OWL ontology for the IFC conceptual data schema and exchange file format for Building Information Model (BIM) data.

Lee, G. et al., 2012. Query Performance of the IFC Model Server Using an Object-Relational Database Approach and a Traditional Relational Database Approach. *Journal of Computing in Civil Engineering*, 28(2), pp.210–222.

Lesniewski, S., 1992. Foundations of the General Theory of Sets (trad. anglaies D. I. Barnett). In S. J. S. et al, ed. *Nijhoff International Philosophy Series*. Polish Scientific Publishers-Kluwer, pp. 129–173.

Li, L. et al., 2014. Modeling and broadening temporal user interest in personalized news recommendation. *Expert Systems with Applications*, 41(7), pp.3168–3177.

Lieberman, 1995. Letizia : An agent that assists web browsing. *IJCAI*, 1, pp.924–929.

Liebich et al., 2013. Industry Foundation Classes IFC4 official release. Available at: http ://www.buildingsmart-tech.org/ifc/IFC4/final/html/index.htm.

Livi, L. & Sadeghian, A., 2016. computational intelligence, and the analysis of non-geometric input spaces, In Granular Computing (March 2016) Vol. 1, Issue 1. *Granular computing*, 1(1), pp.13–20. Available at: https://link.springer.com/article/10.1007/s41066-015-0003-0.

Lohmann, S. et al., 2014. VOWL 2 : User-oriented visualization of ontologies. In *Proceedings of the 19th International Conference on Knowledge Engineering and Knowledge Management (EKAW '14), volume 8876 of LNAI*. pp. 266–281.

Lops, P., De Gemmis, M. & Semeraro, G., 2011. Content-based recommender systems : State of the art and trends. In *Recommender systems handbook*. Springer, pp. 73–105.

Loveland, D.W., 1978. *Automated Theorem Proving: A Logical Basis*, Amsterdam: North-Holland.

Mach, M.A. & Owoc, M.L., 2010. Knowledge granularity and representation of knowledge. Towards knowledge grid. *Intelligent Information Processing V, IFIP - The International Federation for Information Processing*, 340.

Makris, K. et al., 2010. Ontology mapping and SPARQL rewriting for querying federated RDF data sources. *On the Move to Meaningful Internet Systems, OTM 2010*, pp.1108–1117.

Makris, K.E., 2010. Sparql rewriting for query mediation over mapped ontologies.

Mani, I., 1998. A theory of granularity and its application to problems of polysemy and underspecification of meaning. In A. G. Cohn, L. K. Schubert, & S. C. Shapiro, eds. *Principles of knowledge represen-tation and reasoning, Proceedings of the Sixth International Conference (KR98),*. San Mateo: Morgan Kaufmann.

Martín-Guerrero, J.D. et al., 2007. An approach based on the adaptive resonance theory for analysing the viability of recommender systems in a citizen web portal. *Expert Systems with Applications*, 33(3), pp.743–753.

Marwala, T. & Hurwitz, E., 2017. *Artificial intelligence and economic theory: skynet in the market* S. I. Publishing, ed., Springer Cham.

Matheus, C.J., Baclawski, K. & Kokar, M.M., 2006. BaseVISor: A Triples-Based Inference Engine Outfitted to Process RuleML & R-Entailment Rules. In *Proceedings of RuleML 2006*. Athens, Georgia, U.S.A.

McGowan, J., Kushmerick, N. & Smyth, B., 2002. Who do you want to be today ? web personae for personalised information access. In *Adaptive Hypermedia and Adaptive Web-Based Systems*. Springer, pp. 514–517.

Meditskos, G. & Bassiliades, N., 2008. A rule-based object-oriented OWL reasoner. *Knowledge and Data Engineering, IEEE Transactions on*, 20(3), pp.397–410.

Mei, J., Ma, L. & Pan, Y., 2006. Ontology query answering on databases. *The Semantic Web-ISWC 2006*, pp.445–458.

Micarelli, A. et al., 2007. Personalized search on the world wide web. In *The adaptive web*. Springer, pp. 195–230.

Mizzaro, S. & Tasso, C., 2002. Ephemeral and persistent personalization in adaptive information access to scholarly publications on the web. In *Adaptive Hypermedia and Adaptive Web-Based Systems*. Springer, pp. 306–316.

Mladenic, D., 1996. *Personal webwatcher : design and implementation*.

Mockapetris & Dunlap, K.J., 1988. Development of the domain name system.

Morita, M. & Shinoda, Y., 1994. Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. New York: Springer-Verlag, pp. 272–281.

Motik, B., 2006. Reasoning in description logics using resolution and deductive databases.

Motik, B., Patel-Schneider, P.F. & Grau, B.C., 2012. OWL 2 Web Ontology Language Direct Semantics (Second Edition) W3C Recommendation. Available at: https://www.w3.org/TR/owl-direct-semantics/.

Motik, B. & Sattler, U., 2006. A comparison of reasoning techniques for querying large description logic ABoxes. In *Logic for programming, artificial intelligence, and reasoning*. Berlin-Heidelberg: Springer, pp. 227–241.

Motik, B., Sattler, U. & Studer, R., 2005. Query answering for OWL-DL with rules. *Web Semantics : Science, Services and Agents on the World Wide Web*, 3(1), pp.41–60.

Motik, B., Sattler, U. & Studer, R., 2005. Query answering for OWL-DL with rules. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(1), pp.41–60.

Motik, B., Shearer, R. & Horrocks, I., 2009. Hypertableau reasoning for description logics. *Journal of Artificial Intelligence Research*, 36(1), pp.165–228.

Motik, B. & Studer, R., 2005. KAON2–A scalable reasoning tool for the semantic web. In *Proceedings of the 2nd European Semantic Web Conference (ESWC'05), Heraklion, Greece*.

Moukas, A., 1997. Amalthaea information discovery and filtering using a multiagent evolving ecosystem. *Applied Artificial Intelligence*, 11(5), pp.437–457.

Murray, D. & Durrell, K., 2000. Inferring demographic attributes of anonymous internet users. In *LNCS Web Usage Analysis and User Profiling*. Berlin Heidelberg: Springer, pp. 7–20.

Mylonas, P. et al., 2008. Personalized information retrieval based on context and ontological knowledge. *The Knowledge Engineering Review*, 23(01), pp.73–100.

Napoli, A., 1997. *Une introduction aux logiques de descriptions*, INRIA.

Noy, N.F. & McGuiness, D.L., 2001. *Ontology development 101 : a guide to creating your first ontology*, Stantford Medical Informatics.

O'Connor, M.J., Halaschek-Wiener, C. & Musen, M.A., 2010. Mapping Master: a flexible approach for mapping spreadsheets to OWL. *The Semantic Web–ISWC 2010*, pp.194–208.

Oguro, K., 2016. Big data – key to the 4th industrial revolution. *Japan SPOTLIGHT*, pp.24–28. Available at: https://www.rieti.go.jp/en/papers/contribution/oguro/data/07.pdf.

Orłowski, C., Ziółkowski, A. & Czarnecki, A., 2010. Validation of an agent and ontology-based information technology assessment system. *Cybernetics and Systems: An International Journal*, 41(1), pp.62–74.

Pandey, S. et al., 2011. Learning to target : what works for behavioral targeting. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, pp. 1805–1814.

Panetto, H. & Cecil, J., 2013. Information systems for enterprise integration, interoperability and networking : theory and applications. *Enterprise Information Systems*, *7*(1), pp.1–6.

Panetto, H. & Cecil, J., 2013. Information systems for enterprise integration, interoperability and networking: theory and applications. *Enterprise Information Systems*, *7*(1), pp.1–6.

Parent, C. et al., 2008. Modélisation conceptuelle de trajectories. *Revue des Nouvelles Technologies de l'Information*, pp.151–176.

Parent, S., Mobasher, B. & Lytinen, S., 2001. An adaptive agent for web exploration based on concept hierarchies. In *Proceedings of the 9th International Conference on Human Computer Interaction*.

Parsia, B. et al., 2005. Cautiously approaching SWRL. *University of Maryland, Tech. Rep*.

Pasi, G., 2014. Implicit feedback through user-system interactions for defining user models in personalized search. In *The 6th international conference on Intelligent Human Computer Interaction (IHCI)*. Procedia Computer Science, pp. 8–11.

Pauwels, P. & Terkaj, W., 2016. EXPRESS to OWL for construction industry: Towards a recommendable and usable ifcOWL ontology. *Automation in Construction*, 63, pp.100–133.

Pawlak, Z., 1998. Granularity of knowledge, indiscernibility and rough sets. In *Proceedings of the IEEE International Conference on Fuzzy Systems*. pp. 106–110.

Pazzani, M.J., Muramatsu, J. & Billsus, 1996. Syskill & webert : Identifying interesting web sites. *In AAAI/IAAI*, 1, pp.54–61.

Pedrysz, W., 2007. Granular Computing – the emerging paradigm. *Journal of Uncertain Systems*, 1(1), pp.38–61.

Peters, M.A., 2017. Technological Unemployment: Educating for the Fourth Industrial Revolution. *Journal of Self-Governance and Management Economics*, *5*(1), pp.25–33.

Phillips, J. & Phillips, P.P., 2015. High-impact: human capital strategy: addressing the 12 major challenges today's organizations face. *New York: American Management Association*.

Prud'hommeaux, E. & Seaborne, A., 2008. SPARQL query language for RDF - W3C Recommendation 15 January 2008.

R. Fielding, T.B.-L. UC Irvine J. Gettys J. Mogul H. Frystyk L. Masinter P. Leach, 1999. Hypertext Transfer Protocol - HTTP/1.1. Available at: https://tools.ietf.org/html/rfc2616.

Rao, A.S. & George, M.P., 1995. BDI agents: From theory to practice. In *In Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95)*. pp. 312–319.

Rich, E., 1979. User modeling via stereotypes*. *Cognitive science*, 3(4), pp.329–354.

Rifaieh, R., Arara, A. & Benharkat, A.N., 2005. MurO: A Multi-representation Ontology as a Foundation of Enterprise Information Systems. In G. Das & V. P. Gulati, eds. *Intelligent Information Technology: 7th International Conference on Information Technology, CIT 2004, Hyderabad, India, December 20-23, 2004. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 292–301.

Rigas, E., Meditskos, G. & Bassiliades, N., 2012. SWRL2COOL: Object-Oriented Transformation of SWRL in the CLIPS Production Rule Engine. In *SETN*. pp. 49–56.

Rogozov, Y.I. et al., 2013. Data Models based on the Information Granulation theory. In *Proceedings of V International Conference "System Analysis and Information Technologies" SAIT-2013*. Krasnoyarsk, pp. 24–33.

Román, P.E. & Velásquez, J.D., 2014. A neurology-inspired model of web usage. *Neurocomputing*, 131, pp.300–311.

Rosaci, D. & Sarnè, G.M., 2014. Multi-agent technology and ontologies to support personalization in b2c e-commerce. *Electronic Commerce Research and Applications*, 13(1), pp.13–23.

Roxin, A., 2009. *Protocole de découverte sensible au contexte pour les services Web sémantiques*. Université de Technologie de Belfort-Montbéliard (UTBM).

Russell, S. & Norvig, P., 2009. *Artificial Intelligence - A modern approach*. 3rd ed., Egnlewood Cliffs: Prentice-Hall.

Sakagami, H. & Kamba, T., 1997. Learning personal preferences on online newspaper articles from user behaviors. *Computer Networks and ISDN Systems*, 29(8), pp.1447–1455.

Sarraipa, J. et al., 2007. Annotation for enterprise information management traceability. In *ASME 2007 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers, pp. 893–899.

Schevers, H. & Drogemuller, R., 2005. Converting the industry foundation classes to the web ontology language. In *Semantics, Knowledge and Grid, 2005. SKG'05. First International Conference on*. IEEE, pp. 73–73.

Sela, M. et al., 2015. Personalizing news content : An experimental study. *Journal of the Association for Information Science and Technology*, 66(1), pp.1–12.

Shvaiko, P. & Euzenat, J., 2013. Ontology matching: state of the art and future challenges. *Knowledge and Data Engineering, IEEE Transactions on*, 25(1), pp.158–176.

Sieg, A., Mobasher, B. & Burke, R., 2004. Inferring user's information context from user profiles and concept hierarchies. In *Classification, Clustering, and Data Mining Applications*. Springer, pp. 563–573.

Sieg, A., Mobasher, B. & Burke, R., 2007. Web search personalization with ontological user profiles. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, pp. 525–534.

Simons, P., 1987. *Parts: a study in ontology*, Oxford University Press.

Sirin, E. et al., 2007. Pellet: A practical OWL-DL reasoner. *Software Engineering and the Semantic Web*, 5(2), pp.51–53.

Skillen, K. et al., 2014. Ontological user modelling and semantic rule-based reasoning for personalisation of help-on-demand services in pervasive environments. *Future Generation Computer Systems*, 34, pp.97–109.

Smullyan, R.M., 1968. *First Order Logic*, Springer-Verlag.

Sowa, J.F., 1999. Knowledge representation: logical, philosophical, and computational foundations.

Stan, J. et al., 2008. A user profile ontology for situation-aware social networking. In *In 3rd Workshop on Artificial Intelligence Techniques for Ambient Intelligence (AITAmI2008)*.

Stoimenov, L. et al., 2015. Enterprise integration solution for power supply company based on GeoNis interoperability framework. *Data & Knowledge Engineering*.

Stoimenov, L., 2009. Mediation and Ontology-Based Framework for Interoperability. In V. E. Ferraggine, J. H. Doorn, & L. C. Rivero, eds. *Handbook of Research on Innovations in Database Technologies and Applications : Current and Future Trends*. Hershey, PA, USA: IGI Global, pp. 491–507.

Studer, R., Benjamins, V.R. & Fensel, D., 1998. Knowledge engineering: principles and methods. *Data & knowledge engineering*, 25(1), pp.161–197.

T. Berners-Lee, L.M. R. Fielding, 2005. Uniform Resource Identifier (URI): Generic Syntax. Available at: https://www.ietf.org/rfc/rfc3986.txt.

Teevan, J., Dumais, S.T. & Horvitz, E., 2005. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 449–456.

Toch, E., Wang, Y. & Cranor, L.F., 2012. Personalization and privacy : a survey of privacy risks and remedies in personalization-based systems. *User Modeling and User-Adapted Interaction*, 22(1-2), pp.203–220.

Tools, S., 2015. *STEP developer tools reference express-g language overview*, Available at: http://www.steptools.com/support/stdev_docs/devtools/expgoverview.html.

Vallet, D. et al., 2007. Personalized content retrieval in context using ontological knowledge. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(3), pp.3336–346.

Varzi, A., 1996. Parts, Wholes, and Part-Whole Relations : The Prospects of Mereotopology. *Data and Knowledge Engineering*, 20(3), pp.259–286.

Varzi, A.C., 1993. On the boundary between Mereology and Topology. In R. Casati, B. Smith, & G. White, eds. *Proceedings of the 16th International Wittgenstein Symposium Philosophy and the Cognitive Sciences*. Kirchberg am Wechsel, Austria, pp. 423–442.

Vernadat, F. B., 1996. *Enterprise modeling and integration: principles and applications*,

Vernadat, F.B., 2002. Enterprise modeling and integration (EMI): Current status and research perspectives. *Annual Reviews in Control*, 26(1), pp.15–25.

Vildjiounaite, E. et al., 2007. Context-dependent user modelling for smart homes. In *User Modeling*. pp. 345–349.

Vogel, D.S., Asparouhov, O. & Scheffer, T., 2007. Scalable lookahead linear regression trees. In ACM, ed. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '07)*. New York, NY, USA, pp. 757–764.

Volk, R., Stengel, J. & Schultmann, F., 2014. Building Information Modeling (BIM) for existing buildings—Literature review and future needs. *Automation in construction*, 38, pp.109–127.

Vu, T. et al., 2015. Temporal latent topic user profiles for search personalisation. In *Advances in Information Retrieval*. pp. 605–616.

Wærn, A., 2004. User involvement in automatic filtering : An experimental study. *User Modeling and User-Adapted Interaction*, 14(2-3), pp.201–237.

Weichbroth, P. & Owoc, M., 2011. A framework for web usage mining based on multi-agent and expert system an application to web server log files. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, (206), pp.139–151.

Woolridge, M. & Wooldridge, M.J., 2001. *Introduction to Multiagent Systems*, New York, NY, USA: JohnWiley & Sons, Inc.

Xing, B. & Marwala, T., 2018. *Smart maintenance for human–robot interaction: an intelligent search algorithmic perspective*, Springer International Publishing.

Xu, W. & Yu, J., 2017. A novel approach to information fusion in multi-source datasets: A granular computing viewpoint. *Information Sciences*, 378, pp.410–423. Available at: https://doi.org/10.1016/j.ins.2016.04.009.

Yaqoob, I. et al., 2016. Big data: From beginning to future. *International Journal of Information Management*, 36(6 Part B), pp.1231–1247. Available at: http://www.sciencedirect.com/science/article/pii/S0268401216304753.

Zack, M.H., 1999. Managing codified knowledge. *Sloan Management Review, vol*, 40(4), pp.45–58.

Zadeh, L., 1979. Fuzzy sets and information granularity. In Y. R. Ragade R Gupta N, ed. *Advances in Fuzzy Set Theory and Applications*. Amsterdam: North-Holland, pp. 3–18.

Zhang, L. & Issa, R.R., 2011. Development of IFC-based construction industry ontology for information retrieval from IFC Models. In *Proceedings of the 2011 Eg-Ice Workshop*. The Netherlands: University of Twente.

5 Liste des figures

6 Index des tables

7 Index des acronymes

AEC Architecture, Engineering & Construction

AFNOR Association Française pour la NORmalisation

BIM Building Information Modelling

bSDD buildingSmart Data Dictionary

bSI buildingSmart International (avant International Alliance of Interoperability (IAI))

CEM Classical Exetensional Mereology

CEN Comité Européen de Normalisation

COBie Construction Operations Building Information Exchange

CPS Cyber Physical System

CPU Central Processing Unit

CSTB Centre Scientifique et Technique du Bâtiment

CWA Closed World Assumption

DC Dublin Core

DCI Dublin Core Initiative

DIKW Data, Information, Knowledge, Wisdom

DPS Distributed Problem Solving

DSP Demand Side Platform

DL Description logics (Logique de description)

EI Enterprise Interoperability

EII Enterprise Integration and Interoperability

EIS Enterprise Information Systems (Système d'Informations d'Entreprise)

FOAF Friend Of A Friend

FOL First Order Logic

FLOPS Floating Operations Per Second

GIS Geo-spatial Information Systems

GPU Grpahical Processing Unit

GUID Globally Unique Identifier

HTTP HyperText Transfer Protocol

ICDD Information Container for Data Drop

IDM Information Delivery Manual

IETF Internet Engineering Task Force

IFC Industry Foundation Classes

IoT Internet of Things

ISO International Standardization Organization

KNN K-Nearest-Nieghbor (K plus proches voisins)

LD Linked Data

LoD Level of Detail

LDWG Linked Data Working Group

LP Logique propositionnelle

MAS Multi-Agent System

MDA Model Driven Architecture

MDI Model Driven Interoperability

MVD Model View Definition

N3 Notation3

NIBS National Institute of Building Sciences

NNF Normal Negative Form (Forme Négative Normale)

OGC Open Geo-spatial Consortium

OWA Open World Assumption

OWL Web Ontology language

PSD Property Set Definitions

RDF Resource Description Framework

RDFS RDF Schema

RI Révolution Industrielle

RIF Rule Interchange Format

SI Système d'Informations

SIG Système d'Informations géographiques

SHACL Shapes Constraint Language

SOAP Simple Object Access Protocol

SPARQL Simple Protocol And RDF Query Language

SPFF STEP Physical File Format

SPIN SPARQL Inference Notation

SSP Supply Side Platform

STEP Standard for the Exchange of Product model data

SKOS Simple Knowledge Organization System

SWRL Semantic Web Rule Language

Turtle Terse RDF Triple Language

UNA Unique Name Assumption

W3C World Wide Web Consortium

WSDL Web Service Description Language

WWW World Wide Web (W3C)

XML eXtensible Markup Language

XSD eXtensible Schema Definition language

8 Index des termes et des noms propres