Approches sémantiques pour la prédiction de présence d'amiante dans les bâtiments : une approche probabiliste et une approche à base de règles

Thamer Mecharnia

▶ **To cite this version:**

Thamer Mecharnia. Approches sémantiques pour la prédiction de présence d'amiante dans les bâtiments : une approche probabiliste et une approche à base de règles. Intelligence artificielle [cs.AI]. Université Paris-Saclay, 2022. Français. NNT : 2022UPASG036 . tel-03676831

HAL Id: tel-03676831

https://theses.hal.science/tel-03676831v1

Submitted on 24 May 2022

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

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

université PARIS-SACLAY

Approches sémantiques pour la prédiction de présence d'amiante dans les bâtiments : une approche probabiliste et une approche à base de règles

Semantic approaches for predicting the presence of asbestos in buildings : a probabilistic approach and a rule-based approach

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580 Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat : Informatique / Informatique appliquée
Graduate School : Informatique et sciences du numérique, Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche Université Paris-Saclay, CNRS, Laboratoire interdisciplinaire des sciences du numérique, 91405, Orsay, France, sous la co-direction de Nathalie Pernelle, Professeure, la co-direction de Fayçal Hamdi, MCF-HDR et l'encadrement de Lydia Chibout, Docteur.

Thèse soutenue à Paris-Saclay, le 14 Avril 2022, par

Thamer Mecharnia

Composition du jury

Sylvie Despres Professeure, Université Sorbonne Paris Nord (LIM-MICS)	Présidente
Catherine Faron-Zucker MCF-HDR, Université de Nice Sophia Antipolis (I3S)	Rapporteur
Nathalie Hernandez Professeure, Université de Toulouse-Jean Jaures (IRIT)	Rapporteur
Alain Denise Professeur, Université Paris Saclay (LISN)	Examinateur
Bernd Amann Professeur, Sorbonne Université (LIP6)	Examinateur
Nathalie Pernelle Professeure, Université Sorbonne Paris Nord (LIPN)	Directrice de thèse

THESE DE DOCTORAT
NNT : 2022UPASG036

À mes parents et ma sœur

Remerciements

Je voudrais tout d'abord exprimer ma sincère gratitude à mes encadrants Nathalie, Lydia et Fayçal pour m'avoir donné l'opportunité de préparer mon doctorat sous leur direction pendant les trois dernières années. La disponibilité de Nathalie à offrir des conseils, un soutien et des conseils en cas de besoin est quelque chose que j'apprécie vraiment et que je n'oublierai jamais. Son expérience et son intervention aux bons moments, ont massivement contribué à la forme actuelle de ce manuscrit et à la réussite de ce projet. Au CSTB et au Cnam, l'énergie positive, la patience, l'incroyable attention aux détails et l'égalité de traitement que j'ai reçue de Lydia et Fayçal est quelque chose que je n'oublierai jamais. Ils ont consacré une grande partie de leur temps personnel à me guider tout au long de ce voyage, et je leur en serai toujours reconnaissant. Aussi, ayant partagé de nombreux moments ensemble en dehors de notre environnement de travail habituel, j'ai eu la chance de découvrir leurs personnalités joyeuses et leur grand sens de l'humour. Je remercie également Céline Rouveirol pour ses précieux conseils son aide et j'en serai toujours reconnaissant pour ses contributions.

Je remercie les membres de mon comité de thèse, Alain Denise, Bernd Amann, Catherine Faron-Zucker, Nathalie Hernandez et Sylvie Despres pour avoir consacré de leur temps à la lecture du manuscrit et pour leur perspicacité commentaires et remarques.

Je tiens à remercier tous mes collègues du LISN et de l'équipe LaHDAK, de CSTB et de la direction DA2E, et de Cnam de m'avoir soutenu et de m'avoir fait sentir comme faisant partie de leur équipe dès le premier instant. En particulier, ma chaleureuse gratitude va à l'ingénieur recherche et expertise de CSTB Marion Thibault et l'ancien expert Emmanuel Baumont pour leur aide. Je suis reconnaissant pour toutes vos contributions à cette thèse.

Ma chaleureuse gratitude va à mes camarades et mes amis de laboratoire et de DA2E qui ont rendu mon voyage amusant et agréable. Être entouré de très bons amis a rendu ma vie à Paris encore plus belle. Je remercie Fateh, Haitham, Hao, Mouchira, Fatiha, Frédéric et Fu pour les grands moments que nous avons passés au cours de ces trois années.

Ma plus chaleureuse gratitude va à mes parents et ma sœur bien-aimés. Votre amour inconditionnel et vos sacrifices sans fin ont rendu tout cela possible et m'ont amené là où je suis aujourd'hui. Maman, sachant que tu es à mes côtés pendant tous mes meilleurs et pires moments, a rendu ce voyage incroyablement facile. Merci pour votre soutien sans fin. Merci pour vos prières et d'être toujours à mes côtés.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Le Centre Scientifique et Technique du Bâtiment CSTB [1] a été créé en 1947 après la guerre de 1939-1945 afin de superviser la reconstruction des bâtiments en France et archiver la procédure de construction ainsi que les produits et les matériaux utilisés. À cette époque, l'amiante [2] (ou Asbestos en anglais) est utilisé de manière intensive dans les produits utilisés dans les bâtiments (par exemple les tressées ou tissées, les liants, les fibrociments ...). Cette fibre minérale possède en effet des propriétés d'isolation thermique et acoustique, un caractère incombustible, une absence d'usure et une résistance à la traction (matériaux de friction comme les freins et les embrayages) et aux attaques chimiques (sauf avec les acides) [55]. En raison de ses qualités, de nombreux pays, dont la France, ont largement utilisé des produits incluant des fibres d'amiantes que ce soit dans la construction des usines, des immeubles, des établissements scolaires, ou encore des hôpitaux. Cependant, l'amiante a causé la mort de dizaines de milliers de personnes, car ce matériau émet des particules et poussières dangereuses qui sont la cause de nombreuses maladies comme l'asbestose ou fibrose pulmonaire, le cancer des poumons ou le mésothéliome (c.-à-d. tumeur provoquée par la prolifération désordonnée d'un mésothélium) [49]. Grâce à l'évolution de la médecine et grâce aux recherches menées sur les dangers de l'amiante, la production, l'importation et la commercialisation d'amiante sont interdites depuis le premier janvier 1997 en France. La France rejoint ainsi les 7 autres pays européens (Allemagne, Italie, Danemark, Suède, Pays-Bas, Norvège et Suisse) à avoir interdit ce matériau. Néanmoins, il en reste des millions de tonnes disséminés dans les bâtiments construits avant 1997. Le repérage de parties amiantées est donc d'importance que ce soit pour réaliser des travaux de mise en conformité ou pour envisager le recyclage des éléments du bâtiment (p. ex. fenêtre, plancher, porte ...) dans le cadre de l'économie circulaire. Dans le cadre du PRDA [3], le CSTB (Centre Scientifique et Technique du Bâtiment) a été sollicité pour développer un outil en ligne d'aide à l'identification de matériaux contenant potentiellement de l'amiante dans

1. http://www.cstb.fr/fr/
2. https://fr.wikipedia.org/wiki/Amiante
3. Plan de recherche et de développement amiante lancé par la Direction de l'Habitat, de l'Urbanisme et des Paysages (DHUP), rattachée à la Direction Générale de l'Amiante, de l'Habitat et de la Nature (Ministre du Logement et de l'Habitat Durable)

les bâtiments afin de guider l'opérateur dans la préparation de son programme de suivi (Projet ORIGAMI). Le but du projet ORIGAMI est de créer un outil fournissant une aide au repérage de matériaux amiantés dans les bâtiments. Cet outil a pour objectif d'orienter l'opérateur de repérage et de prioriser les bâtiments sur lesquels effectuer un diagnostic. L'objectif du repérage amiante avant travaux est d'éviter d'exposer les artisans, ouvriers ou usagers du bâtiment à l'amiante. En effet, la réalisation de travaux sur des matériaux amiantés est susceptible de libérer des fibres d'amiante. L'inhalation accidentelle de ces fibres représente un risque sanitaire important puisque celles-ci sont considérées comme cancérigènes. Cet outil ne se substituera en aucun cas au repérage des matériaux et produits contenant de l'amiante réalisé par un professionnel conformément à la norme *NF X46-020*[4].

Problématique et Objectifs

Compte tenu du grand nombre de bâtiments à étudier et à analyser, les experts du CSTB ont besoin d'un outil pour les aider à prioriser les parties de bâtiments sur lesquelles ils doivent réaliser des diagnostics. Cet outil doit permettre de représenter et d'analyser les descriptions des bâtiments disponibles dans les archives du CSTB pour associer à un produit utilisé dans un bâtiment une connaissance sur la présence ou l'absence d'amiante. Cependant, les données disponibles présentent certaines particularités qui rendent leur exploitation difficile :

— Le CSTB conserve différents types de documents dont les projets homologués qui sont des documents qui décrivent des types de bâtiments construits éventuellement en plusieurs exemplaires (c.-à-d. des immeubles, des écoles, des hôpitaux, etc.), et les diagnostics qui décrivent les résultats positifs ou négatifs des prélèvements déjà effectués sur des parties de bâtiments particuliers. Les descriptions des bâtiments sont plus ou moins structurées dans un format PDF, XML ou Excel. Les projets homologués et les diagnostics sont des documents textuels stockés en PDF ou en CSV qui comportent des données tabulaires décrivant les bâtiments. Cependant, la structure de ces tableaux diffère d'un fournisseur à l'autre, d'où la nécessité d'utiliser des techniques de NLP pour détecter les termes du domaine.

— Il n'existe pas encore des vocabulaires normalisés du domaine du bâtiment qui décrivent ses constituants. Or, afin d'intégrer les données issues de différentes sources et de les exploiter, il est nécessaire de disposer d'un vocabulaire du domaine potentiellement réutilisable pour d'autres problématiques liées au bâtiment (plomb, réutilisation et économie circulaire, environnement et énergie).

— Dans les documents du CSTB, nous ne connaissons que les types du produit utilisés lors de la construction, mais nous ne connaissons pas la référence du produit commercialisé réellement utilisé. Par exemple, nous savons que le produit utilisé est de type "enduit", mais nous ne savons pas de quel enduit précis il s'agit exactement (p. ex., la référence de l'enduit commercialisé appelé "FILGUM"). Cette incomplétude des don-

4. https://www.boutique.afnor.org/norme/nf-x46-020/reperage-amiante-reperage-des-materiaux-et-produits-contenant-de-l-amiante-dans-les-immeubles-batis-mission-et-methodologie/article/867769/fa186482

nées ne permet pas d'utiliser directement les connaissances sur la présence d'amiante dans la référence FILGUM pour en déduire la présence d'amiante dans un bâtiment.

— Comme les méfaits de l'amiante ont été rendus publics à partir des années 1959 (année des premières publications des séries de cas et suggestions d'une relation possible entre l'amiante et le mésothéliome [49]), son utilisation a commencé à décroître et, par conséquent, sa présence dans les produits commercialisés est devenue étroitement liée à l'année de construction du bâtiment. Ainsi, il est essentiel de prendre en compte ces aspects temporels dans le traitement des données liées à l'amiante.

— Certains organismes comme l'INRS (Institut National de Recherche et de Sécurité [5]) et l'Andeva (Association Nationale de Défense des Victimes de l'Amiante [6]) ont publié des listes de produits commercialisés qui indiquent pour chaque produit les périodes durant lesquelles ils ont été identifiés comme amiantés. Cependant, ces informations ne sont pas connues pour toutes les périodes (aucune information n'est disponible pour certains produits et certains intervalles de temps) et il existe des informations conflictuelles entre les informations de ces deux ressources. Par exemple, le même produit commercialisé peut être décrit comme étant amianté dans la ressource publiée par l'INRS et non amianté durant le même intervalle de temps pour l'Andeva (exemple : "ISOCOL COLLE PU 403" a été amianté jusqu'à 1980 pour l'Andeva, mais il est resté encore amianté jusqu'à 1983 pour l'INRS).

Pour répondre à ces problèmes, nous proposons dans cette thèse des approches qui exploitent les données du CSTB sur les bâtiments, les connaissances des experts et les ressources externes pour prioriser les diagnostics à effectuer en prenant en compte des données temporelles, incomplètes et parfois contradictoires. L'objectif est non seulement de prédire la présence d'amiante dans les produits des bâtiments, mais d'expliquer cette prédiction en fournissant des règles interprétables qui ont permis de l'établir.

La prédiction de présence d'amiante dans les produits est réalisée via un processus de classification qui associe à chaque produit une classe (amianté ou non amianté). La classification, qui est un problème classique en apprentissage, permet de classifier, à l'aide d'approches supervisées et non supervisées, des données représentées sous forme de vecteurs de caractéristiques. Dans le domaine des graphes de connaissances différentes approches ont abordé ce problème en s'intéressant à la découverte de règles logiques du premier ordre qui sont apprises à partir de données relationnelles étiquetées [19, 47] ou à partir d'exemples et de contre-exemples générés à partir d'hypothèses. Cependant, les données sont généralement incomplètes et les contre-exemples ne sont pas toujours disponibles (notamment à cause de l'hypothèse du monde ouvert, sur laquelle sont fondées les ontologies, qui suppose qu'un fait qui n'est pas dans un graphe de connaissances n'est pas nécessairement faux). De plus du problème de l'incomplétude des données, la sémantique de l'ontologie peut être exploitée par l'expert lorsqu'elle est disponible pour obtenir des descriptions plus détaillées sur les composants. Certaines approches non supervisées

5. https://www.inrs.fr/
6. https://andeva.fr/

visent à découvrir des patrons de graphes dans des graphes RDF volumineux sans prendre en compte l'ontologie [27, 43]. Les problèmes de ces approches par rapport aux caractéristiques de nos données et de notre domaine sont liés au fait que nous traitons de données temporelles où nous voulons disposer dans les règles de prédicats intégrés avec des constantes numériques qui apparaissent une seule fois dans les règles. À cause de la propriété de fermeture de règle (voir Section 3.3.2), ces approches ne permettent pas de trouver des règles satisfaisant ce critère temporel.

L'objectif est donc de définir des approches de fouille de règles, fortement guidé par le contexte et les relations entre les composants du bâtiment (les siblings [7] et la hiérarchie de type représentés par le modèle partie-tout [6]).

Contributions

Dans le cadre de cette thèse, nous avons modélisé et peuplé une ontologie de domaine pour représenter les bâtiments et leurs éléments constitutifs et nous avons proposé deux approches basées sur cette ontologie qui permettent de générer un ensemble de règles concluant sur la présence ou l'absence d'amiante dans un élément constitutif du bâtiment.

La première approche se base sur des ressources externes au CSTB qui décrivent les produits commercialisés pour calculer une probabilité de présence d'amiante. Tandis que la deuxième utilise le graphe de connaissances construit pour apprendre des règles de raisonnement dans le but de classer les produits avec des raisonneurs qui utilisent ces règles découvertes.

Plus précisément, les contributions de ce travail de thèse sont les suivantes :

1. La construction, l'enrichissement et le peuplement d'une ontologie ASBESTOS qui décrit en détail les bâtiments, ses composants et ses propriétés.
2. La création d'un module qui peuple l'ontologie de base par des informations sur l'amiante (extraites depuis des ressources internes au CSTB et des ressources externes), et qui suit les standards de la time ontologie [8]. Ce module a été validé par l'expert et peut être réutilisable à d'autres problématiques au sein du CSTB (module amiante remplacé par module plomb, autres ressources externes ...).
3. Une première approche hybride, probabiliste et sémantique, qui formalise les données temporelles et éventuellement contradictoires issues des ressources externes et les utilise pour estimer une probabilité de présence d'amiante dans les éléments de bâtiments (c.-à-d. produits, localisations, structure et bâtiment).
4. Une deuxième approche, appelée CRA-Miner, qui se base sur les exemples positifs et négatifs décrits dans les diagnostics pour générer un ensemble de règles permettant de classer les produits.

7. Les siblings sont les individus dans le niveau n dans le graphe qui possèdent le même type de relation avec le même individu dans le niveau $n-1$
8. Time ontology : `https://www.w3.org/TR/owl-time/`

5. Une implémentation des approches proposées dans cette thèse et la création d'un outil qui enrichit et peuple l'ontologie et qui utilise le raisonneur Pellet [9] pour classer les produits.
6. Un ensemble d'expérimentations réalisées pour (1) valider les résultats de nos approches et pour (2) les comparer avec d'autres méthodes d'apprentissage sur les graphes de connaissances.

Ontologie Asbestos, enrichissement et peuplement : Comme il n'existe pas une ontologie du domaine du bâtiment et de l'amiante, nous nous sommes basés sur les données du CSTB et les ressources externes pour la mise en place d'une ontologie appelée Asbestos qui représente les connaissances sur les bâtiments, ses composants et ses propriétés d'amiante. Cette ontologie est enrichie et peuplée avec un ensemble varié de ressources. Durant l'enrichissement et le peuplement de l'ontologie, nous avons extrait les données et les connaissances depuis les ressources tabulaires ou textuelles. L'ontologie Asbestos est divisée en plusieurs modules pour faciliter sa maintenance ainsi que sa réutilisation dans d'autres domaines.

Approche hybride de détection des produits amiantés basée sur des ressources externes : La première approche se base sur l'ontologie ASBESTOS et sur les ressources externes fournies par l'INRS et l'ANDEVA décrivant la présence d'amiante dans les produits commercialisés. Dans un premier temps, les connaissances temporelles décrites dans ces ressources sont fusionnées et structurées selon l'ontologie ASBESTOS. Le processus de fusion suit une approche pessimiste en cas d'information contradictoire sur le même produit commercialisé : le fait que l'une des sources indique que le produit comme amianté suffit à le considérer ainsi. Dans un second temps, pour une année de construction du bâtiment donnée, une probabilité est calculée pour chaque classe de produit. Les parts de marchés des différents produits commercialisés n'étant pas connues, le calcul proposé suppose que chaque produit commercialisé est équiprobable. Enfin, comme ces ressources ne listent que les produits commercialisés ayant été amiantés à un moment de leur commercialisation, nous avons utilisé un ensemble de diagnostics pour ajuster les probabilités calculées et limiter l'impact de ce biais.

Les expérimentations effectuées sur un jeu de données du CSTB montrent que la précision de cette première approche hybride est bonne, mais que la couverture peut être améliorée, car les ressources externes sont incomplètes.

Approche logique de classification des produits (CRA-Miner) : Cette deuxième approche se base sur les diagnostics archivés au sein du CSTB pour découvrir des règles permettant de classifier les produits amiantés ou non-amiantés. Comme nous souhaitons obtenir toutes les règles logiques permettant de conclure sur la présence ou l'absence d'amiante, nous nous sommes inspirés des techniques de fouille de règles définies pour les graphes de connaissances pour définir une approche *top-down* permettant de générer toutes les règles dont le *head coverage* et la confiance sont supérieures à un seuil. La particularité de notre approche est d'être capable de prendre en compte une partie de la sémantique de l'ontologie, les caractéristiques de l'évolution du nombre de produits

9. Pellet : `https://www.w3.org/2001/sw/wiki/Pellet`

amiantés au fur et à mesure du temps ainsi que des heuristiques propres aux relations parties-tout. Ces dernières permettent de limiter l'espace de recherche tout en s'autorisant la découverte de règles complexes prenant en compte le contexte dans lequel le produit se positionne dans le bâtiment (c.-à-d. complexe en termes de prédicats apparaissant en prémisse de la règle). Ce type d'approche ne garantissant pas qu'une seule règle puisse être appliquée sur un produit, elle peut aboutir à la génération de conclusions contradictoires pour certains produits. Comme il est important dans ce cadre de limiter les faux négatifs, nous avons défini une approche pessimiste permettant de favoriser la déduction de présence d'amiante.

CRA-Miner a été testé sur un jeu de données issu du CSTB et les résultats ont été comparés à une approche naïve (*baseline*), à notre approche hybride basée sur les ressources externes, à l'approche de fouille de règles AMIE3 ainsi qu'à une approche relationnelle de type *divide-and-conqueer* (i.e. TILDE). Les résultats montrent que la couverture de CRA-Miner est meilleure que celle obtenue par la première approche, la *baseline* et TILDE, et que la précision est significativement meilleure que celle obtenue par AMIE3.

Plan du Manuscrit

Ce manuscrit comporte 6 chapitres.

Chapitre 2 : Définitions et concepts de base

Ce premier chapitre introduit les concepts et les définitions de base qui seront nécessaires pour présenter l'état de l'art et les approches proposées dans les chapitres suivants.

Chapitre 3 : Revue de l'état de l'art

Ce chapitre est un état de l'art correspondant aux travaux que nous avons réalisés. Nous avons présenté les différentes stratégies et approches de fouilles de règles et de liage de données, et nous avons positionné nos deux approches par rapport à l'existant.

Chapitre 4 : Construction d'une ontologie de l'amiante

Nous présentons dans ce chapitre notre ontologie ASBESTOS et la méthodologie suivie pour la construire. Nous présentons également les données et les ressources dont nous disposons, ainsi que la méthode suivie pour extraire ces données depuis les ressources documentaires. Nous montrons aussi la procédure d'enrichissement et de peuplement de l'ontologie avec les données.

Chapitre 5 : Calcul de probabilité de présence d'amiante dans les parties de bâtiments basé sur des ressources externes

Ce chapitre est dédié à notre première approche de classification de produits, que nous avons appelée "approche hybride", et qui est une approche semi-supervisée pour la prédiction de l'amiante basée sur notre ontologie ASBESTOS et les ressources externes en utilisant des données temporelles, incomplètes et contradictoires. Nous montrons aussi dans ce chapitre comment nous traitons les données conflictuelles dans les ressources. Nous expliquons également la procédure de calcul de la probabilité de présence d'amiante et la classification des données du CSTB selon les probabilités calculées.

Chapitre 6 : Découverte de règles contextuelles de prédictions à partir de données de diagnostics amiante

Ce chapitre est dédié à la deuxième approche de fouille de règles d'apprentissage pour la classification des produits. Cette approche, appelée CAR-Miner, utilise des données étiquetées pour apprendre des règles de classification en s'inspirant des techniques de Programmation Logique Inductive (PLI). Les règles sont apprises en prenant en compte l'aspect temporel des données et un contexte complexe et riche en relation qui est défini par l'expert.

Enfin nous concluons et nous donnons quelques perspectives.

Chapitre 2

Définitions et concepts de base

2.1 Introduction

Nous introduisons dans ce chapitre les concepts et les définitions de base qui seront nécessaires pour présenter l'état de l'art et les approches proposées dans les chapitres suivants.

Nous allons commencer tout d'abord par présenter la notion de graphe de données RDF dans la section 2.2. Puis nous définissons la notion d'ontologie, de graphe de connaissances et ces différents éléments (Section 2.3). Dans la section suivante, nous présentons les trois familles du langage d'ontologie Web OWL (Section 2.5). Puis nous définissons les règles SWRL et les raisonneurs dans les sections 2.6 et 2.7. Finalement, dans la section 2.8 nous présentons l'hypothèse du monde ouvert.

2.2 Graphe de données RDF

Dans le contexte du Web sémantique, les données peuvent être décrites au format RDF [1] (Resource Description Framework), qui est un modèle de représentation de données normalisé par le W3C. Le modèle RDF représente les données sous forme de triplets qui peuvent être notés sous la forme suivante : $< sujet, predicat, objet >$. Le sujet représente la ressource décrite par le triplet, le prédicat représente une propriété et l'objet représente un littéral ou une ressource. Une ressource peut représenter n'importe quel objet abstrait ou concret comme, par exemple, une personne, une organisation, une propriété, une image ou un fichier. Elle est référencée par une chaîne de caractères appelée URI (Uniform Resource Identifier) telle que http ://dbpedia.org/resource/Philomena, ou par une IRI (Internationalized Resource Identifier) qui est une chaîne de caractères représentée en Unicode. Les littéraux peuvent être associés à des types de données qui définissent l'étendue des valeurs possibles, tels que les chaînes de caractères, les nombres ou encore les dates. De plus, il est possible de modéliser des nœuds blancs qui repré-

1. Grammaire de la syntaxe rdf : `https://www.w3.org/TR/rdf-syntax-grammar/`

sentent des ressources dont on connaît l'existence, mais pour lesquelles on ne connaît pas l'identifiant.
Par exemple, le triplet $< b1, region, "ile\ de\ France" >$, représente le fait que la ressource $b1$ est située dans la région "île de France" qui est représentée ici par une valeur littérale, tandis que le triplet $< b1, hasStructure, s1 >$, représente le fait que $b1$ est associé à une structure $s1$ qui est une URI qui référence également une ressource.

Un graphe de données RDF peut être défini de la façon suivante :

Définition 1 (graphe de données RDF) *Soient U un ensemble d'URI, L un ensemble de littéraux et B un ensemble de noeuds blancs, un graphe de données RDF est un ensemble de triplets $< s, p, o >$ tels que $s \in U \cup B$, $p \in U$ et $o \in U \cup B \cup L$.*

Un graphe de données RDF peut également être représenté par un graphe labellisé et dirigé où un nœud représente une ressource, un littéral ou un nœud vide tandis qu'un arc représente une propriété. Ces graphes peuvent être sérialisés pour les rendre partageables et publiables sur le Web sous différents formats comme RDF/XML [2], Turtle [3] (Terse RDF Triple Language), ou encore N-Triples [4] qui est un sous-ensemble de Turtle et qui, comme Turtle, est plus facile à interpréter, plus lisible que RDF/XML.

Pour extraire ou manipuler les informations d'un jeu de données RDF, des requêtes peuvent être définies en SPARQL [5] (SPARQL Protocol and RDF Query Language) qui est un langage de requête et un protocole d'accès normalisé par le W3C et adapté à la structure des graphes RDF.

Un graphe de données décrit donc des objets d'intérêt et des connexions entre ces objets. Par exemple, un graphe peut comporter des nœuds décrivant un livre, les auteurs de ce livre, les éditeurs de ce livre dans différents pays, etc. Chaque nœud pouvant avoir des propriétés littérales telles que le nom et l'âge d'un auteur, ou encore les langues officielles du pays. Un utilisateur ou une application peut alors parcourir le ou les graphes pour recueillir des informations sur tous les livres.

2.3 Ontologie et graphe de connaissances

Dans le cadre du Web sémantique, le partage et la mise à jour des connaissances ont toujours constitué un problème compte tenu de l'hétérogénéité des données et des vocabulaires utilisés. La modélisation et l'exploitation des ontologies qui définissent le vocabulaire d'un domaine facilitent l'interprétabilité des données et l'interopérabilité des applications. De nombreuses définitions du terme ontologie ont été proposées. Gruber (1993) [23] en a proposé la définition suivante « Une ontologie est une spécification explicite d'une conceptualisation ». En d'autres termes, une ontologie est un modèle de domaine qui est décrit explicitement. Lassila et McGuinness (2001) [28] définissent

2. Grammaire de la syntaxe rdf : `https://www.w3.org/TR/rdf-syntax-grammar/`
3. Turtle : `https://www.w3.org/TR/turtle/`
4. N-Triples : `https://www.w3.org/TR/n-triples/`
5. Langage de requête SPARQL 1.1 : `https://www.w3.org/TR/sparql11-query/`

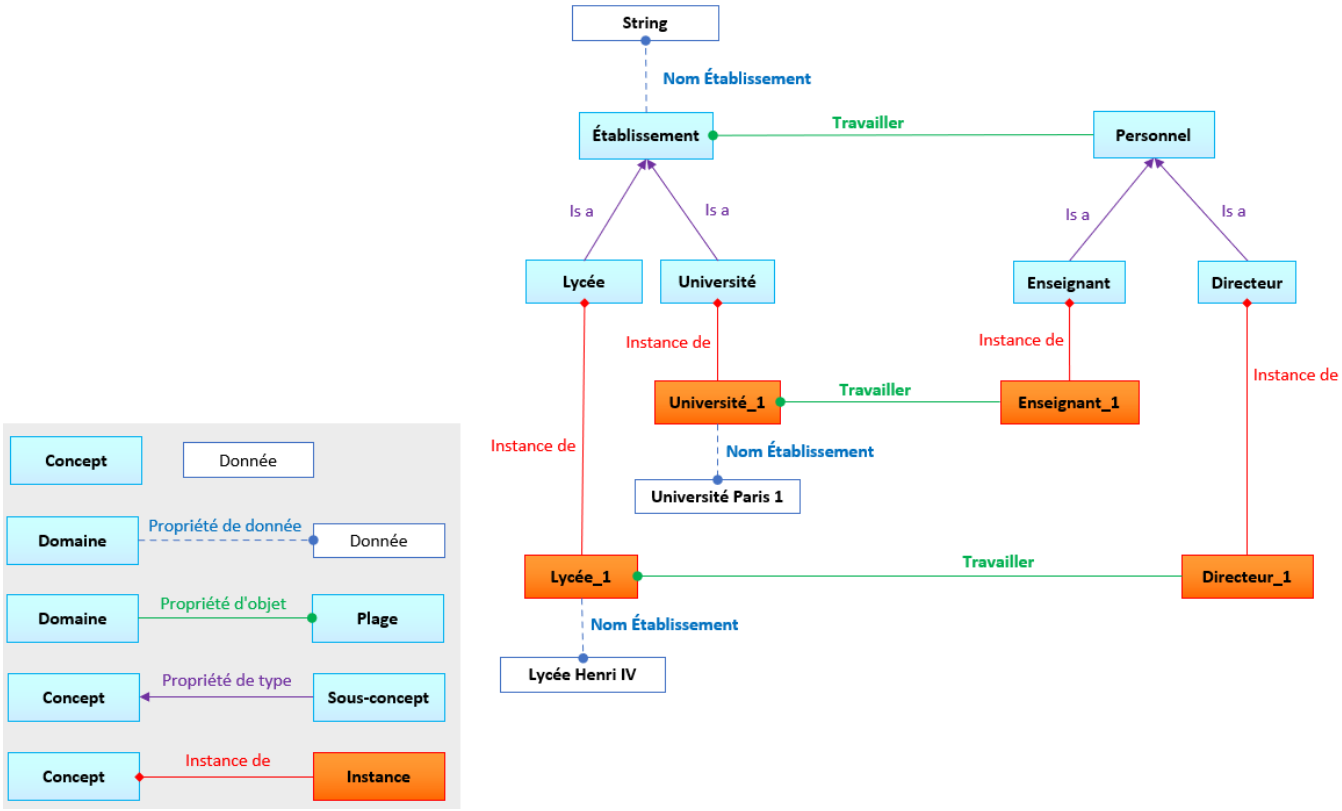


FIGURE 2.1 – Ensemble de concepts et de propriétés de l'ontologie de l'éducation

les propriétés obligatoires pour qu'une telle spécification soit mise en oeuvre : un vocabulaire fini et contrôlé, une interprétation non ambiguë des classes et des relations entre les termes et des relations de spécialisation entre classes.

Une ontologie peut être représentée par un ensemble de concepts, de propriétés, et d'axiomes qui représentent un domaine tel que la santé, les bâtiments, la culture, l'éducation, ou encore l'agronomie. Par exemple, l'ontologie présentée en Figure 2.1) décrit des établissements d'enseignement et permet de représenter le vocabulaire nécessaire pour représenter un lycée ou une université ainsi que le personnel qui y travaille.

Une ontologie comporte principalement les éléments suivants :

— Les concepts ou classes qui représentent les types des individus.

— Les attributs ou propriétés de type de donnée (data property) qui représentent des propriétés dont les valeurs sont littérales, numériques, temporelles, booléennes, etc.

— Les relations ou propriétés d'objet (object property) qui représentent les liens entre individus

— Les axiomes qui décrivent les liens de spécialisation entre classes, les disjonctions entre classes, les propriétés fonctionnelles, les cardinalités des propriétés ...

Une ontologie qui inclut les données et donc les descriptions des instances de classe peut également être appelée ontologie peuplée, ou Graphe de connaissances (GC). Formellement, un graphe de connaissances peut

être défini comme suit [13] :

Définition 2 (Graphe de connaissances) *Un graphe de connaissance est défini par un tuple* $o = \langle C, I, R, T, V, \sqsubseteq, \perp, \mathcal{A}, \in, = \rangle$ *tel que :*

C *est un ensemble de concepts.*

I *est un ensemble d'individus.*

R *est un ensemble de relations.*

T *est un ensemble de types de données.*

V *est ensemble de valeurs (*C, I, R, T, V *étant deux à deux disjoints).*

$\sqsubseteq$ *est une relation dans* $(C \times C) \cup (R \times R) \cup (T \times T)$ *appelée spécialisation (ou subsomption).*

$\perp$ *est une relation dans* $(C \times C) \cup (R \times R) \cup (T \times T)$ *appelée exclusion (ou disjonction).*

$\mathcal{A}$ *est un ensemble d'axiomes (autres que la spécialisation et l'exclusion .*

$\in$ *est une relation sur* $(I \times C) \cup (V \times T)$ *appelé instanciation.*

$=$ *est une relation sur* $I \times R \times (I \cup V)$ *appelé affectation.*

D'après la définition 2, les relations d'affectation ($=$) sont définies pour les propriétés d'objet et les propriétés de données, c.-à-d. si nous prenons la relation $I \times R \times I$ nous nous focalisons sur l'affectation des propriétés d'objet, tandis que la relation $I \times R \times V$ définis l'affectation des propriétés de données.

Une ontologie représente donc un certain vocabulaire utilisé pour décrire de façon explicite et consensuelle une certaine réalité. Elle est utilisée comme outil de communication entre humains, entre humain et machine ou entre machines. Elle peut être utilisée pour représenter les connaissances propres à un domaine particulier (e.g. Gene ontology, GeoNames), des connaissances plus génériques (e.g. Yago) ou des connaissances de haut niveau ou trans-domaines (e.g. DOLCE, OWL-Time). Comme une ontologie est facilement partageable sur le Web, elle peut être réutilisée et intégrée dans une autre ontologie via une procédure d'alignement qui vise à détecter et découvrir des correspondances entre deux ontologies pour découvrir par exemple que le concept "produit" dans une première ontologie est équivalent au concept "matériau" dans une deuxième ontologie. La réutilisation et la mise à jour des ontologies peuvent être facilitées par la création d'ontologies modulaires, qui permettent de faciliter leur validation, leur maintenance et leur visualisation [11]. L'édition d'une ontologie peut être réalisée via différents éditeurs d'ontologie comme Protégé, SemanticWorks, SWOOP, ou Tedi. Certains éditeurs permettent de visualiser la forme graphique d'une ontologie.

Enfin, une ontologie permet d'effectuer un raisonnement automatique sur les données (Section 2.7) et de rendre ainsi explicites des données qui peuvent être déduites en s'appuyant sur la sémantique de l'ontologie. Cela peut être réalisé en vue de détecter des erreurs dans la formalisation, dans les données ou de produire un graphe dit saturé pour lequel on a jouté toutes les inférences possibles.

La formalisation d'une ontologie est effectuée à l'aide des langages RDFS ou OWL (Langage d'ontologie Web)

qui se basent sur le standard RDF et la syntaxe XML.

2.4 Le langage RDFS

Le langage RDFS (Resource Description Framework Schema) est un langage standard du W3C qui étend le vocabulaire RDF avec un ensemble de termes permettant de décrire le schéma d'un graphe RDF (https ://www.w3.org/TR/rdf-schema). RDFS décrit des propriétés et des classes de ressources RDF. Les classes et les propriétés sont identifiées par des URI ou IRI et sont décrites à l'aide des ressources RDF Schema rdfs :Class et rdfs :Property.

RDFS permet de spécifier les contraintes ontologiques suivantes sur les classes et les propriétés :

— Les contraintes de spécialisation de classe représentées par des triplets de la forme, $< A\ rdfs : subClassOf\ B >$ qui spécifie que la classe A est une sous-classe de la classe B, c'est-à-dire que chaque instance i de A est une instance de B,

— Les contraintes de spécialisation de propriétés représentées par des triplets de la forme, $< p_1\ rdfs : subPropertyOf\ p_2 >$ qui spécifie que la propriété p_1 est plus spécifique que la propriété p_2 et donc que toute paire de ressources liées par p_1 est également liée par p_2 :

— Les contraintes de domaine représentées par des triplets de la forme, <p rdfs :domain A>) qui exprime que le sujet d'une propriété p est une instance de la classe A,

— Les contraintes de co-domaine représentées par des triplets de la forme, $< p\ rdfs : range\ B >$ qui exprime que l'objet d'une propriété p est une instance de la classe B.

2.5 Langages d'ontologie Web OWL et OWL 2

Le Langage d'Ontologie Web (OWL) est également un standard du W3C, il s'appuie sur RDF et RDFS pour permettre la construction et le partage des ontologies. Il existe trois sous-langages de OWL [54] : OWL-Lite, OWL-DL et OWL-Full.

— OWL-Lite supporte les fonctionnalités qui utilisent une hiérarchie de classification et les contraintes simples, il permet l'utilisation de propriétés de type de données (*datatype* properties) et l'accès aux valeurs de données.

— OWL-DL inclut toutes les constructions du langage OWL en permettant l'utilisation des fonctions de calcul. il se base principalement sur la logique de description DL.

— OWL-Full base sur une sémantique différente de OWL-Lite ou OWL-DL. il a été conçu pour préserver une certaine compatibilité avec RDFS. Cependant, OWL-Full est indécidable, donc aucun raisonneur n'est capable d'effectuer un raisonnement complet pour le OWL-Full.

OWL s'appuie sur XML Schema (xsd) pour la liste des types de données comme les chaînes de caractères, les

entiers et les dates. La nouvelle version de OWL appelée OWL 2 [6] améliore considérablement les types de données et ajoute un nouvel ensemble de constructeurs pour les types de données (Built-in Datatypes). Ces constructeurs aident à valider les données de l'ontologie comme xsd :pattern qui exige que la donnée doive respecter une certaine expression régulière. Ou des restrictions sur la longueur de la donnée acceptée. Par exemple, dans OWL nous pouvons déclarer que chaque personne a un âge, qui est un entier, mais nous ne pouvons pas restreindre la plage de ce type de données pour dire que les adultes ont un âge supérieur à 18 ans. OWL 2 fournit de nouveaux constructeurs pour les types de données comme xsd :minInclusive qui permet de donner une valeur minimale de l'âge d'une personne.

Une ontologie OWL 2 consiste en : un ensemble d'axiomes de classe qui spécifient les relations logiques entre les classes, qui constitue la boîte terminologique (TBox) ; un ensemble d'axiomes de propriété pour spécifier les relations logiques entre les propriétés, qui constitue une boîte de rôle (RBox) ; et une collection d'assertions décrivant des individus, qui constitue une Assertion Box (ABox).

Les classes représentent la description formelle d'un ensemble d'objets (ou individu) qui partage les mêmes propriétés et les mêmes caractéristiques. Les objets appartenant à la même classe sont considérés comme des objets de même type.

Les propriétés dans OWL 2 sont catégorisées en deux types : des propriétés de données (data properties) et des propriétés d'objet (object properties). Les propriétés de données représentent des relations binaires entre des objets et des valeurs de données (extraites des types de données XML Schema comme : int, string, dateTime, etc). Les propriétés d'objet représentent des relations binaires entre des objets.

La Figure 2.2 représente un aperçu du langage OWL 2 [7], montrant ses principaux éléments constitutifs et leurs relations les uns avec les autres. L'ellipse au centre représente la notion abstraite d'ontologie, qui peut être considérée soit comme une structure abstraite, soit comme un graphe RDF. En haut se trouvent diverses syntaxes concrètes qui peuvent être utilisées pour sérialiser et échanger des ontologies. En bas se trouvent les deux spécifications sémantiques qui définissent la signification des ontologies OWL 2.

OWL 2 utilise la logique de description $\mathcal{SROIQ}(\mathcal{D})$, cette logique de description correspond à l'ensemble des constructeurs disponibles dans OWL 2 prenant en charge la négation des atomes, l'intersection de concepts, les expressions de classe complexes en combinant des opérateurs mathématiques tels que les relations de sous-classe, l'équivalence, la conjonction, la disjonction et la négation, union de concepts, hiérarchie de propriétés d'objet (sous-propriétés), inclusion de propriétés complexes, disjonction de propriétés, propriétés inverses, propriétés de type de données, valeurs de données et types de données. OWL 2 DL [17] est une réalisation de $\mathcal{SROIQ}(\mathcal{D})$ qui permet la définition des classes, des propriétés et des individus.

6. OWL 2 : `https://www.w3.org/TR/owl2-overview/`
7. La structure de OWL2 : `https://www.w3.org/TR/owl2-overview/#Overview`

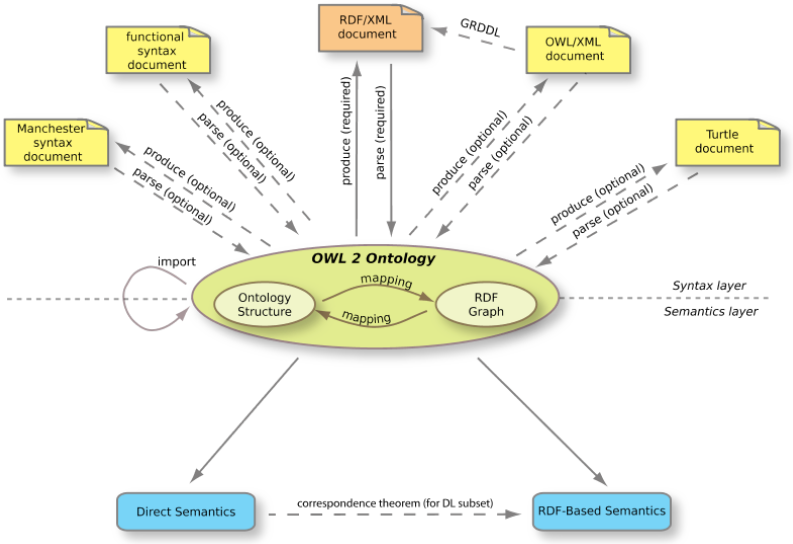


FIGURE 2.2 – La structure de OWL 2

2.6 Règle SWRL

Les langages OWL ou OWL2 ne permettent pas d'exprimer toutes les relations entre les individus. Un exemple connu est que l'on ne peut pas exprimer la relation *enfant de parents mariés*, car il n'y a aucun moyen dans OWL ou OWL2 d'exprimer la relation entre des individus avec lesquels un individu a des relations.

L'expressivité de OWL peut être étendue en ajoutant des règles SWRL (Semantic Web Rule Language) à une ontology. SWRL [8] est un langage de règles pour le Web sémantique. Il résulte d'une combinaison entre le langage OWL-DL et le langage RuleML (Rule Markup Language) [24]. Une règle SWRL est une clause de Horn de la forme $Body \rightarrow Head$ qui exprime que la conclusion de la règle ($Head$) doit être vraie lorsque la prémisse de la règle ($Body$) est satisfaite.

Plus précisément, dans une règle SWRL notée $\overrightarrow{B} \rightarrow \overrightarrow{H}$, $\overrightarrow{B}$ et $\overrightarrow{H}$ sont des conjonctions d'atomes définis comme suit : $\overrightarrow{B} = \{B_1, ..., B_n\}$ et $\overrightarrow{H} = \{H_1, ..., H_m\}$. Aussi, la forme générale d'une règle SWRL est : $B_1, ..., B_n \rightarrow H_1, ..., H_m$. Les atomes dans le corps et la tête de la règle sont des prédicats unaires que l'on peut noter $C(t)$ ou des prédicats binaires $R(t_1, t_2)$. Dans le cas des prédicats unaires, C représente la classe du terme t. Pour les prédicats binaires, R représente la relation entre les deux termes t_1 et t_2. Dans le cas des prédicats binaires, les termes t_1 et t_2 peuvent être des variables ou des constantes (numériques, littérales, dates, etc.).
Par exemple, la règle SWRL suivante exprime le fait que si une personne est âgée de plus de 18 ans alors elle est considérée comme adulte :

personne(?P), age(?P, ?A), greaterThanOrEqual(?A, 18) $\rightarrow$ *adulte(?P)*

Les trois premiers prédicats constituent le corps de la règle (la prémisse) et le dernier constitue la tête de la règle

8. SWRL : `https://www.w3.org/Submission/SWRL/`

	Pellet	Ontop	Mastro	jcel	HermiT	FaCT++	ELK
Subsomption	Oui	Non	Non	Oui	Non	Non	Non
Satisfaisabilité	Oui	Non	Oui	Oui	Oui	Oui	Oui
Classification	Oui	Non	Oui	Oui	Oui	Oui	Oui
Récupération d'instance	Oui	Non	Non	Oui	Non	Non	Non
Réponse à une requête conjonctive	Oui	Oui	Oui	Non	Non	Non	Non

TABLE 2.1 – Comparaison des raisonneurs en fonction des services de raisonnement supportés

(la conclusion). Dans cette règle, nous avons utilisé un prédicat intégré (greaterThanOrEqual) qui est prédéfini par SWRL afin d'effectuer une comparaison avec une valeur numérique.

Les règles SWRL peuvent être utilisées par les raisonneurs (voir section 2.7) pour inférer de nouvelles instances de classes ou de propriétés dans l'ontologie. La restriction aux règles DL-Safe, qui s'appliquent uniquement aux individus nommés, permet de rester décidable.

2.7 Raisonneurs

La possibilité de réaliser des inférences à partir d'un graphe de connaissances est essentielle et représente l'un des avantages majeurs de l'utilisation des ontologies et des règles SWRL qui peuvent leur être associées. Les raisonneurs sont des moteurs d'inférence qui rendent cette tâche possible. Il existe différents raisonneurs qui ont chacun leurs propriétés [9]. Ils se différencient par les environnements dans lesquels ils peuvent être intégrés, par les services de raisonnement supportés et par les méthodologies .

Certains raisonneurs peuvent être utilisés à partir de l'environnement "Protégé" [10] qui permet de créer et manipuler des bases ontologiques simples ou complexes. C'est le cas de ELK, FaCT++, HermiT, jcel, ontop et Pellet. D'autres sont aussi supportés par la bibliothèque owlready2 [11] de python qui permet la gestion des ontologies (Pellet et HermiT). Pellet, HermiT et FaCT++ sont des raisonneurs qui prennent en charge les spécifications de OWL 2. Cependant, seul Pellet peut utiliser les prédicats intégrés de SWRL tels que greaterThan, lessThan, ou encore lessThanOrEqual, permettant de comparer des données numériques ou de type date. La table 2.1 montre les services de raisonnement supportés par chacun de ces raisonneurs.

Dans cette thèse, nous avons choisi d'utiliser *Pellet* pour sa compatibilité avec OWL (voir section 2.5) et pour sa capacité à intégrer les règles SWRL dans le moteur d'inférences OWL-DL fondé sur les algorithmes des tableaux sémantiques. Pellet permet d'utiliser les prédicats intégrés de SWRL qui permettent de comparer des constantes de divers types tels que les entiers, les dates, etc. De plus, son intégration dans un prototype développé en python est facile (c.-à-d. il est disponible dans la bibliothèque "owlready2" [12] de python).

9. Liste des raisonneurs existants : `http://owl.cs.manchester.ac.uk/tools/list-of-reasoners/`
10. `https://protege.stanford.edu/about.php`
11. La bibliothèque Owlready2 : `https://pypi.org/project/Owlready2/`
12. Documentation d'Owlready2 : `https://owlready2.readthedocs.io/en/v0.36/`

2.8 Hypothèse du monde ouvert

Les graphes de connaissances sont de plus en plus nombreux et volumineux, mais les données ne sont pas nécessairement complètes. L'hypothèse du monde ouvert (Open World Assumption - OWA) [20] indique que ce qui n'est pas connu pour être vrai n'est pas nécessairement faux. En pratique, cela signifie que des conclusions ne peuvent être tirées qu'à partir de faits et d'axiomes explicitement énoncés. C'est le contraire de l'hypothèse du monde fermé (Closed World Assumption - CWA), qui indique que tout fait vrai est connu pour être vrai. L'hypothèse du monde fermé permet à une application de déduire, de son manque de connaissance d'un fait, que ce fait est faux et d'inférer ce que qui découle de ce fait considéré comme faux. Les langages du Web sémantique tels que OWL font l'hypothèse du monde ouvert. Un raisonneur ne peut pas déduire qu'un fait est faux parce qu'il est absent.

Cette hypothèse a également des conséquences sur les approches de découvertes de connaissances telles que l'apprentissage de règles à partir de graphes de connaissances. Ces graphes ne contiennent pas toujours d'exemples négatifs et certaines approches utilisent des hypothèses particulières sur les données. L'approche de [20] a montré comment découvrir des règles à partir de GC malgré l'absence de contre-exemples explicites. L'hypothèse clé est l'hypothèse de complétude partielle (Partial Completeness Assumption - PCA). Cela permet à [20] de considérer des contre-exemples de règles, même dans le cadre de l'OWA.

Chapitre 3

Revue de l'état de l'art

3.1 Introduction

Dans ce chapitre, nous présentons un état de l'art correspondant aux travaux que nous avons réalisés. Nos travaux se basent sur des données du CSTB décrivant des bâtiments et des diagnostics pour construire un graphe de connaissances et utiliser ces données pour prédire la présence d'amiante dans les produits utilisés dans des parties de bâtiments non diagnostiqués. Cette prédiction peut soit se baser sur des connaissances externes décrivant les listes de références de produits amiantés, soit sur des règles de classification induites à partir des diagnostics déjà décrits dans le graphe.

Aussi, nous nous focalisons dans ce chapitre sur les approches de liage de données et sur les approches de découvertes de règles. Les approches de liage de données peuvent permettre, en créant des liens d'identités, de compléter un graphe de connaissance pour lequel certains faits sont manquants. Ces approches se basent sur la comparaison des descriptions RDF de deux objets pour décider si ces descriptions réfèrent bien au même objet du monde réel. Dans le contexte des graphes de connaissances (GC), la fouille de règles peut également être utilisée pour enrichir les graphes et permettre de prédire de nouveaux liens ou types, ou permettre de détecter des triplets RDF erronés. Motivées par le besoin de passage à l'échelle, la plupart des approches récentes de prédiction de liens ou de types sont basées sur des méthodes d'apprentissage profond et de plongement de graphes (*graph embeddings*) qui permettent de traduire des vecteurs de grande dimension en espaces de dimension relativement faible [44]. Néanmoins, d'autres applications pour lesquelles des règles interprétables sont nécessaires pour comprendre et maintenir une certaine connaissance du domaine sont toujours intéressées par la découverte de règles logiques.

Dans cet état de l'art, nous allons tout d'abord présenter les stratégies de liage de données basées sur des règles. Nous présenterons ensuite les différentes stratégies d'induction de règles définies en Programmation Logique Inductive (PLI) (Section 3.3) avant de présenter les stratégies d'induction de règles logiques définies pour les

graphes de connaissance, et terminerons par la définition des biais de langage qui a une grande importance pour limiter l'espace de recherches et ainsi permettre son exploration en un temps raisonnable, et pour obtenir des résultats qui soient d'intérêts pour les utilisateurs. Puis nous allons discuter l'évolution des approches qui s'intéressent à l'apprentissage de règles, en expliquant les procédures et les types de données traitées par ces approches (Section 3.3.2). Nous présentons également dans cette partie, les mesures de qualités utilisées par ces approches pour évaluer les règles découvertes et les caractéristiques des règles à trouver.

3.2 Approches de liage de données

Le liage de donnée consiste à déterminer si deux descriptions réfèrent au même objet du monde réel. Il est souvent utilisé dans le Web sémantique pour établir des liens entre les données. LinkingOpenData [1] est un des projets de liage de données des graphes RDF où les liens entre les données sont explicitement déclarés en utilisant le constructeur owl :sameAs. Différents types de méthodes permettent de lier les données. Certaines plateformes se basent sur des règles de liage déclarées par un expert du domaine [56]. Cependant, il n'est pas toujours facile pour un expert de spécifier toutes les propriétés ou combinaisons de propriétés qui peuvent être discriminantes pour un ensemble d'individus. Le numéro de sécurité sociale est une propriété clairement discriminante pour des personnes, mais il est plus difficile de spécifier que le nom,la date et le lieu de naissance suffisent à s'assurer qu'il s'agit bien de la même personne.

Aussi, certaines approches sont supervisées et se basent sur des échantillons de données liées et de données distinctes. D'autres approches se basent des sur des jeux de données RDF [53, 41] pour lesquelles on peut supposer que l'*Unique Name Assumption* est respectée. Dans ce cadre, il s'agit de déterminer toutes les propriétés qui suffisent à distinguer un individu d'un autre individu sans avoir besoin d'exemples positifs. Les règles apprises peuvent être plus ou moins expressives. Les approches telles que [52, 53] s'intéressent uniquement à des éléments de descriptions de longueur 1, qui peuvent comporter des constantes. Ainsi [53] permet de découvrir la règle de liage conditionnelle suivante : $Ville(X) \wedge Ville(Y) \wedge Dpartement(X, \#Aisne) \wedge Dpartement(Y, \#Aisne) \wedge Nom(X, N) \wedge Nom(Y, N) \rightarrow sameAS(X, Y)$ Cette règle signifie que deux villes qui partagent le même nom sont identiques si elles sont situées dans le département de l'Aisne.

D'autres approches se basent sur la recherche de motifs de graphes plus complexes qui représentent des expressions référentielles. Une telle approche permet de découvrir par exemple que Mozart est la seule personne qui est un musicien reconnu et qui est né à Salzbourg en 1756. Pour limiter l'espace de recherche, [41] ne s'intéresse qu'aux combinaisons de propriétés instanciées apparaissant dans des non-clefs.

Quelle que soit l'approche de liage proposée, il est nécessaire de disposer d'informations communes dans les deux jeux de données qui soient suffisamment discriminantes pour décider que les deux IRI correspondent à deux

1. https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

individus identiques.

3.3 Stratégies d'induction de règles

L'induction de règle peut être effectuée en suivant différents types de stratégie et en choisissant celle qui convient le mieux aux besoins de l'ontologue, à la composition et aux différents types des prédicats qui apparaissent dans la règle.

Dans cette section, nous présentons les principes de la programmation logique inductive (PLI) suivis par les différents types de stratégies. Par la suite, nous décrirons les différentes approches d'induction de règles. Enfin, nous aborderons les différentes contraintes posées par le biais de langage qui guident les stratégies et bornent l'espace de recherche.

3.3.1 Programmation Logique Inductive PLI

Les approches existantes de découverte de règles de la logique du premier ordre dans les graphes de connaissances sont généralement inspirées des méthodes développées en PLI (Programmation Logique Inductive) [38]. L'objectif de la PLI est d'apprendre d'une hypothèse, c.-à-d. un ensemble de règles de la forme "IF prémisse ALORS conclusion", à partir d'un ensemble de données appelées exemples. Les exemples fournis pour un apprentissage supervisé correspondent à des couples constitués de la description relationnelle d'un exemple et d'une étiquette correspondant à l'une des classes à apprendre. Lorsque l'on ne dispose que de deux classes (ensemble $E+$ et $E-$ d'exemples et de contre-exemples des instances de classe), il s'agit d'un problème d'apprentissage de concept. Le problème peut alors être défini comme suit :

"*L'apprentissage de concepts est la construction d'une description générale d'une classe d'objets à partir d'un ensemble d'exemples positifs et d'exemples négatifs.*" [36]

Une autre définition formelle est donnée par [10], à partir d'un langage de concepts L_C, un langage d'exemples L_e, les couvertures ou relations d'appartenance $\in_C$ qui spécifie comment L_C se rapporte à L_e, et un ensemble d'exemples E d'un concept cible inconnu $t \in L_C$. Chaque exemple est de la forme $(e, Classe)$ où $e \in L_e$, et $Classe$ est $vrai$ ou $faux$. Les exemples $(e, vrai)$ sont des exemples positifs, tandis que les exemples $(e, faux)$ sont négatifs. L'objectif de l'apprentissage des concepts est alors de trouver une hypothèse $H \in L_C$ qui couvre tous les exemples positifs (dans ce cas H est complet) et aucun des exemples négatifs (dans ce cas H est cohérent).

L'espace de recherche correspond à un ensemble d'hypothèses, pour lequel on doit définir un langage d'hypothèses permettant de délimiter la représentation des concepts à apprendre pour réduire l'espace de recherche et garantir une certaine qualité des résultats en évitant au système de considérer des hypothèses inutiles ou difficiles à interpréter. Il s'agit des biais de langage. Il en est de même pour le langage des exemples et celui de la théorie

du domaine. Le parcours de l'espace de recherche est effectué grâce à des opérations de généralisation et/ou de spécialisation basées sur une relation de subsomption. A priori, une hypothèse devrait classifier correctement tous les exemples positifs (complétude) et aucun exemple négatif (correction). Cependant, les exemples d'apprentissage comportent souvent des données erronées ou des exceptions, il est donc parfois difficile de trouver une hypothèse qui soit à la fois complète et correcte. Aussi, la plupart des approches relaxent cette contrainte et essaient de trouver une hypothèse qui couvre de nombreux exemples positifs en couvrant aussi peu d'exemples négatifs que possible. Les algorithmes développés sont soit ascendants ou descendants. Les algorithmes ascendants tels que [40, 50] partent d'un exemple positif et le généralise tant que l'hypothèse ne couvre pas ou peu de négatifs tandis que les algorithmes descendants tels que [2, 47, 40, 25] considèrent une prémisse de règle vide ou correspondant aux hypothèses les plus générales que l'on peut examiner (i.e top(s) de l'espace de recherche) et spécialisent ces hypothèses.

L'apprentissage des règles logiques de prédiction est considéré comme un problème de recherche qui est souvent NP-difficile, il faut donc chercher la solution en optimisant l'espace de recherche sachant que la complexité de l'espace de recherche dépend de la complexité des solutions.

Stratégies de type Separate-and-conquer

Separate-and-Conquer (SAC) et Divide-and-Conquer (DAC) sont des stratégies très populaires pour l'induction de règles logiques. Ces deux stratégies de base sont suivies par différentes approches d'induction des règles sur les graphes de connaissances.

Separate-and-Conquer (SAC) (appelé aussi *covering algorithms* [18]) démarre par la production d'un ensemble de règles en spécialisant à plusieurs reprises une règle plus générale, en exploitant que les exemples positifs, c.-à-d. ces règles découvertes ne couvrent que des exemples positifs. À chaque itération, SAC sélectionne une règle plus spécialisée que si elle couvre un sous-ensemble des exemples positifs et exclut les exemples négatifs. Ceci est répété jusqu'à ce que tous les exemples positifs soient couverts par l'ensemble de règles.

Stratégies de type Divide-and-Conquer

Par ailleurs, Divide-and-Conquer (DAC) [4] produit une hypothèse en divisant une règle trop générale en un ensemble de règles spécialisées qui couvrent des sous-ensembles disjoints d'exemples. Les règles qui couvrent uniquement les exemples positifs sont conservées, tandis que les règles qui couvrent à la fois les exemples positifs et négatifs sont traitées récursivement de la même manière que la première règle générale. Plus précisément, si DAC trouve une règle qui couvre que des exemples positifs, elle l'ajoute au résultat. Si une règle couvre que des exemples négatifs elle l'exclue du résultat. Cependant, si elle trouve une règle R qui couvre des exemples positifs et négatifs, elle la divise en $R_1, ...R_n$ telle que $R_1 \cup ... \cup R_n \Leftrightarrow R$, puis elle fait la même chose pour chaque règle R_i

avec $i \in [1, n]$.

Ces deux stratégies présentent des coûts de calcul différents. Pour SAC, le coût de calcul mesuré en termes de nombre de vérifications pour voir si une règle couvre ou non un exemple augmente de manière quadratique avec la taille de l'ensemble d'exemples, alors qu'il augmente linéairement pour DAC. Cela découle du fait que SAC recherche un espace d'hypothèse plus grand que DAC.

Parmi les approches de type DAC nous trouvons celles qui utilisent les arbres de décision logique du premier ordre FOLDT (First-Order Logical Decision Tree) telles que TDIDT (Top-down induction of decision trees) [46] et son successeur TILDE (Top-down induction of logical decision trees) [2]. Ces dernières se basent sur des arbres de décision dans lesquels les noeuds peuvent partager des variables et impliquer des prédicats numériques comportant des valeurs seuils. Cependant, TDIDT et TILDE n'utilisent pas la sémantique de l'ontologie dans l'exploration de l'espace de recherche.

TILDE (Top-down induction of logical decision trees) [2] se base sur le système de l'induction descendante des arbres de décision TDIDT [46] qui suit la stratégie de DAC. TDIDT utilise les arbres de décision logique du premier ordre FOLDT pour découvrir des hypothèses H (ou des règles) à partir d'un ensemble de classes C, un ensemble d'exemples E et un ensemble de connaissances B, tel que pour tout $e \in E, H \wedge e \wedge B \models c$ et $e \in E, H \wedge e \wedge B \not\models c'$, avec c est la classe de l'exemple e et $c' \in C - \{c\}$. L'hypothèse H est composée de prédicats et de classes qui sont représentés dans les nœuds de FOLDT. Ces nœuds peuvent se retrouvés dans l'arbre T comme des feuilles avec une classe k et dans ce cas ils sont notés $T = leaf(k)$, ou des nœuds internes avec un prédicat $pred$ et deux branches gauche l et droite r, et dans ce cas le nœud est noté $T = inode(pred, l, r)$. Au fur et à mesure que l'exploration de l'arbre avance dans les branches, TILDE enrichit l'hypothèse H par les prédicats trouvés dans les nœuds de la branche en cours d'exploration et qui se termine par la classe localisée dans la feuille (le nœud final de la branche). L'exploration des branches est guidée par un biais de langage déclaratif (voir Section 3.3.1) pour limiter l'espace de recherche.

Comme cité précédemment, TILDE est fondé sur les mêmes techniques que TDIDT et calcule en plus le nombre de tests à considérer à un nœud. Ce nombre est trouvé par l'opérateur de raffinement $\theta - subsomption$, c.-à-d. étant donné deux clauses[2] c_1 et c_2, on dit que $c_1\ \theta - subsume\ c_2$ ssi $\exists \theta : c_1\theta \in c_2$ avec θ est appelé une substitution de variable.

TILDE utilise également un autre opérateur appelé $lookahead$ qui permet d'effectuer plusieurs étapes de raffinement successives à la fois pour découvrir et ajouter plusieurs prédicats au même temps. Par exemple, $lookahead(pred(x, y), r(x))$ spécifie qu'à chaque fois le prédicat $pred$ est ajouté, un raffinement supplémentaire en ajoutant $r(x)$ (avec x la première variable de $pred$) doit être effectué dans la même étape de raffinement. Cette opération permet d'explorer des variables qui peuvent être significatives pour la règle.

TILDE permet aussi l'utilisation des variables numérique à condition qu'elles apparaissent parmi les données,

2. Une clause est une conjonction de prédicats

comme il peut préciser son utilisation dans les modes du biais de langage déclaratif.

Stratégies de type Reconsider-and-Conquer

Dans le but de gérer plus efficacement le problème de la récursivité que DAC en reconsidérant certaines des décisions antérieures, et de permettre une induction plus efficace que SAC en s'accrochant à certaines des décisions. Une autre stratégie hybride Reconsider-and-Conquer (RAC) [5] a été proposée pour combiner et améliorer DAC et SAC.

La première amélioration de SAC apportée par RAC est d'ajouter un ensemble de règles au lieu d'ajouter une règle à la fois. La première règle de cet ensemble est générée comme dans SAC en spécialisant la règle initiale en une règle qui ne couvre que des exemples positifs. Cependant, au lieu de continuer la recherche d'une règle suivante à partir de la nouvelle règle, RAC revient à l'étape précédente pour examiner si une autre étape de spécialisation pourrait être prise afin de couvrir certains des exemples positifs restants, c.-à-d. RAC va chercher d'autre(s) spécialisation(s) possible de la règle initiale pour créer d'autre(s) branche(s). Cette méthode est similaire à DAC sauf que RAC est moins limité que DAC en ce qui concerne les étapes possibles de spécialisation qui peuvent être prises lors du retour en arrière, car les étapes de spécialisation ne sont pas choisies indépendamment par DAC parce que les règles résultantes devraient constituer une division de la règle initiale (la règle spécialisée dans DAC doit être une division de la règle initiale), or RAC peut trouver des branches indépendantes.

Pour continuer l'exploration d'une branche, la condition suivie par RAC est la fraction d'exemples positifs parmi les exemples couverts ne doit jamais diminuer durant l'exploration de la branche (car cela indiquerait que la branche se concentre sur la couverture des exemples négatifs plutôt que positifs).

Néanmoins, il existe deux caractéristiques de RAC : la première serait que chaque règle durant l'exploration d'une branche couvre au moins un exemple positif. Cela ferait en sorte que RAC se comporte de manière très similaire à DAC. La deuxième serait de toujours exiger de RAC de revenir à la règle initiale pour rendre son comportement identique à celui de SAC.

Stratégies de type Generate-and-Test

Le but des stratégies présentées précédemment est de classer chaque individu une seule fois au maximum dont le résultat trouvé comporte une seule règle par classe d'individu ou par concept. Cependant, il existe plusieurs possibilités de combinaisons pour la même règle qui peut être appliquée à la même classe et certaines de ces combinaisons peuvent être plus précises et plus exactes dans certains cas. D'où la stratégie Generate-and-Test (GAT) [29] qui n'utilise pas les mêmes critères d'arrêt que les SAC, DAC et RAC, mais des mesures de qualité sur les règles comme la confiance (Section 3.3.2) ce qui permet de générer un ensemble plus large et plus divers de règles.

GAT utilise aussi les principes de la Programmation Logique Inductive PLI (Section 3.3.1) pour trouver des règles de prédiction. Le but de cette stratégie est d'explorer toutes les règles qui suivent certains critères comme le biais de langage (Section 3.3.1) et une longueur maximale et qui respectent les mesures de qualité.

Cette stratégie comporte deux étapes principales, elle commence par générer des règles satisfaisant des critères prédéfinis (comme le biais de langage), dans le but de garantir que les règles à générer ne sont pas redondantes ni triviales. Puis, pour chaque règle, elle utilise des métriques pour évaluer la qualité des règles. Ces métriques sont les mesures de qualité définies pour les règles.

Biais de langage

Sans aucune limitation sur l'espace de recherche, la découverte des règles coûtera un temps immense pour la recherche. Pour borner la recherche de règles, les approches et les méthodes de découvert de règles ajoutent des restrictions sur l'exploration des données et la génération des hypothèses à tester. Parmi les restrictions nous trouvons celles qui limitent le nombre de variables et le nombre d'entités dans les prédicats de la règle, mais aussi d'autres qui limitent le nombre et le type des prédicats eux-mêmes, ce qui définira la longueur de la règle. Ces restrictions peuvent être classées en deux catégories, des restrictions syntaxiques sur la forme des prédicats dans la règle, et des restrictions sémantiques sur le comportement des règles induites [1].

L'utilisation des restrictions dans les approches qui basent principalement sur PLI nous définis le biais de langage, vu la nécessité de coder ces restrictions pour les utiliser dans les approches, nous trouvons des méthodes qui utilisent déclarations de mode [37] et d'autres qui basent sur les méta-règles (metarules) [7].

La forme de déclarations de mode peut définir quels prédicats peuvent d'être ajoutés à la règle ainsi que le nombre d'apparitions de ces prédicats. Cette forme met aussi des restrictions sur le type de paramètres des prédicats. Cette forme de biais de langage est utilisée par plusieurs approches comme [2] qui utilise le pattern ou le mode de la règle déclaré par le biais de langage pour chercher que les règles qui suivirent ce pattern. Ces patterns sont déclarés dans un fichier qui englobe tous les modes qui définissent le biais de langage, par exemple :

$$rmode(N : conj).$$

où l'opérateur $rmode$ signifie que le système peut ajouter la conjonction (ou le prédicat) $conj$ à la règle N fois, c.-à-d. N représente le nombre d'apparitions de $conj$ dans la règle. Nous trouvons plus de détailles sur les déclarations et les opérateurs utilisés dans [3]. Pour garantir les propriétés des règles (présentés dans la Section 3.3.2) comme la fermeture (Définition 4) et la connexité (Définition 3), ce type de biais utilise trois modes de base :

- $+$ signifie que la variable est unifiée avec une variable déjà présente dans la règle. Par exemple $rmode(conj(+X)).$ signifie que la variable X doit être présente dans un autre prédicat qui est déjà dans la règle.
- $-$ signifie qu'il s'agit d'une nouvelle variable, c.-à-d. aucune unification n'est effectuée avec des variables déjà existantes. Il faut alors la représenter avec un nouveau littéral.

- $\setminus$ signifie qu'une nouvelle variable doit être placée ici, et une contrainte doit être ajoutée à la clause indiquant que la valeur de la variable doit être différente des valeurs de toute autre variable du même type.

Ce Biais contient aussi un mode pour indiquer la possibilité d'utiliser des constantes et les positions de ces constantes parmi les paramètres de prédicats, nous parlons du mode $\#$ qui représente les constantes, par exemple : $rmode(conj(-X, \#))$. signifie que le deuxième paramètre est une constante tandis que le premier est une nouvelle variable.

Ils existent d'autres approches qui n'utilise pas le PLI pour la recherche de règle, alors ils vont trouver des clauses de second ordre, et dans ce cas ils nécessitent un biais de langage plus adapté à ses besoins et qui permettre d'exprimer des règles logiques de second ordre. Pour répondre à leurs besoins, ces approches utilisent les méta-règles. L'avantage des méta-règles est qu'elles permettent de définir l'espace de recherche des systèmes non PLI pour formaliser les règles logiques de second ordre. Contrairement à la première forme de biais de langage (les modes) qui est utilisé pour les systèmes qui se basent sur le PLI, les méta-règles sont elles-mêmes des énoncés logiques.

3.3.2 Stratégies définies pour les graphes de connaissances

De nombreuses approches se sont intéressées à l'apprentissage de règles et de concepts dans les graphes de connaissances et les ontologies. Les règles découvertes permettent en effet de réaliser de la complétion de graphe, de détecter des données erronées, de découvrir des règles facilitant l'intégration de données comme les règles de liage ou des règles de correspondance définies au niveau conceptuel. A la différence des approches classiques définies en PLI, certaines de ces approches peuvent considérer l'incomplétude des données, la sémantique de l'ontologie et donc des mesures de qualité et plus généralement des biais de langage qui peuvent être mieux adaptés aux graphes de connaissance.

Dans cette section, nous allons parler de différentes approches qui focalisent sur l'apprentissage de concept et ceux qui s'intéressent à la fouille de règles d'association. Ces approches évaluent les règles d'association avec différentes mesures de qualité, et cherche seulement des règles avec des caractéristiques précises alors chacune utilise un biais de langage spécifique à leurs besoins.

Apprentissage de concepts représentés en logique de description

Les approches d'apprentissage de concepts telles que DL-Foil [14] ou DL-FOCL [48] permettent d'apprendre des définitions de concepts représentées en logique de description. Ces approches s'appuient sur des stratégies de type *separate-and-conquer* qui permettent de construire une disjonction de solutions partielles, pouvant être spécialisées à l'aide d'opérateurs de raffinement basés sur la subsomption, de façon à couvrir autant d'exemples positifs que possible tout en excluant (presque) tous les exemples négatifs. Cependant, ces approches, si elles

permettent de générer des définitions de concepts dans des logiques de description expressives, ne recherchent pas toutes les solutions partielles, et donc toutes les définitions. De plus, elles ne permettent pas d'utiliser des prédicats instanciés par des constantes ou de rechercher des valeurs seuils (p. ex. $X \leq 17$ pour définir un mineur).

Fouille de règles d'association

Les approches telles que AMIE3 [27] et RuDiK [43] s'intéressent à la découverte d'ensembles de règles exprimées en logique du premier ordre (clauses de Horn) dans des données RDF (voir section 2.2) volumineuses. Ces deux approches sont de type *générer et tester*. Cependant, elles diffèrent sur la nature des règles pouvant être découvertes, sur les mesures de qualité utilisées et sur certaines hypothèses concernant la complétude des données.

AMIE3 débute en considérant une file de règles qui contient initialement tous les atomes de tête possibles, et qui sont donc des règles de longueur 1. Pour chacune de ces règles, si sa confiance (Section 3.3.2) dépasse le seuil et si la règle est fermée (Définition 4) elle va être ajoutée à l'ensemble de résultats. Tant que la longueur d'une règle dans la file d'attente est inférieure à la longueur maximale prédéfinie, un processus de raffinement est appliqué qui étend cette règle pour produire un ensemble de nouvelles règles. Ces nouvelles règles, si elles ne sont ni dupliquées ni élaguées par le seuil de *head coverage* (Section 3.3.2), vont être ajoutées à la file d'attente de règles (qui contient les règles éligibles au processus de raffinement). Ce processus est répété jusqu'à ce que la file d'attente soit vide. La définition d'une longueur maximale qui inclut aussi le prédicat de la tête permet à AMIE3 de mieux contrôler l'espace de recherche.

Le processus de raffinement utilise plusieurs opérateurs pour enrichir une règle avec de nouveaux atomes :

— Ajouter un atome suspendu (Dangling Atom $\mathcal{O}_D$) : cet opérateur ajoute un nouvel atome avec deux variables, tel qu'il partage une de ces variables avec un autre atome qui est déjà dans la règle.

— Ajouter un atome instancié (Instantiated Atom $\mathcal{O}_I$) : cet opérateur ajoute un nouvel atome avec une variable et une entité (constante), sous la condition que sa variable est partagée avec un autre atome qui est déjà dans la règle.

— Ajouter un atome de clôture (Closing Atom $\mathcal{O}_C$) : dans ce cas il ajoute un atome tel que ces deux variables sont partagées avec les autres atomes de la règle.

Pour être conservées, les règles créées par ces opérateurs doivent avoir une meilleure qualité que la règle initiale.

Les opérateurs du processus de raffinement assurent que les règles seront connexes (Définition 3). Cependant ils peuvent trouver des règles non fermées (Définition 4) à cause de la longueur maximale de la règle qui peut ne pas laisser la possibilité au processus de raffinement d'ajouter un atome de clôture. Il est alors nécessaire de vérifier la fermeture de la règle avant de l'ajouter à l'ensemble de résultats.

AMIE3 permet de découvrir toutes les règles de type $\overrightarrow{B} \rightarrow h(x, y)$ telles que $\overrightarrow{B}$ est une conjonction de prédicats, telles que la confiance ou la PCA-confiance est maximale, pouvant éventuellement comporter des constantes, et

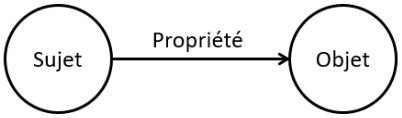


FIGURE 3.1 – Représentation d'un triplet dans un graphe orienté

telles que les règles respectent le seuil de *head coverage*.

Pour disposer de contre-exemples, AMIE3 [27] se base sur l'hypothèse de complétude partielle (Partial Completeness Assumption - PCA) qui suppose que si l'on connait un y pour un certain $pred$ et x d'un triplet $pred(x, y)$, alors on connait l'ensemble des y qui satisfont ce triplet (avec les mêmes $pred$ et x). Cela signifie que si on suppose que lorsqu'un objet est représenté pour une entité et une propriété spécifique, tous les objets sont représentés (c.-à-d. les autres étant considérés comme contre-exemples, voir section 2.8).

Cependant, Amie3 ne permet pas de découvrir des règles comportant une comparaison avec des constantes dans les prédicats de la règle. RuDiK fournit cette possibilité et permet de découvrir des règles telles que :

$$Marrie(x, y) \wedge Habite(x, z) \Rightarrow Habite(y, z)$$

En outre, il permet aussi de découvrir des règles comportant des négations (i.e. not).

$$DateDeNaissance(a, d_1) \wedge DateDeNaissance(b, d_2) \wedge d_1 > d_2 \Rightarrow notFils(a, b)$$

Le but de RuDiK [43] (Rule Discovery in Knowledge Bases) est de trouver des règles représentées en logique du premier ordre (clause de horn) qui couvrent la majorité d'exemples positifs et le moins possible d'exemples négatifs. Les règles découvertes peuvent être utilisées dans deux cas : (1) pour générer de nouveaux faits, elles sont appelées dans ce cas des règles positives, (2) ou pour identifier les faits erronés, elles sont considérées dans ce cas comme des règles négatives. Pour illustrer les règles négatives, [43] donne l'exemple suivant qui exprime que si une personne b est née avant une autre personne a, elle ne peut pas être son fils :

$$DateDeNaissance(a, d_1) \wedge DateDeNaissance(b, d_2) \wedge d_1 > d_2 \wedge Fils(a, b) \Rightarrow \perp$$

En exécutant cette règle comme une requête sur des faits de $Fils$, RuDiK peut identifier des triplets erronés.

Dans cet exemple, RuDiK permet l'utilisation des constantes qu'il s'agisse d'URI décrivant des individus ou de littéraux dans les atomes de la règle. Il permet en plus d'exploiter les opérateurs de comparaison suivants : $\{<, >, \leq, \geq, \neq\}$ sous la condition que les deux membres de la comparaison soient des littéraux. Sauf $\neq$ qui peut être utilisé entre deux variables.

Pour générer les règles, RuDiK suppose que le graphe de connaissances contient tous les exemples positifs et négatifs nécessaires. Ils considèrent le graphe comme un graphe orienté où les sujets et les objets sont représentés par des nœuds, et les propriétés sont représentées par des arêtes orientées du sujet vers l'objet (Figure 3.1).

Le corps de la règle est généré en suivant un chemin dans ce graphe orienté. Sachant que RuDiK utilise un biais de langage pour trouver des règles sûres (Définition 5), fermées (Définition 4) et connexes (Définition 3), le chemin couvert par le corps doit couvrir les variables de la tète au moins une fois (pour valider la sûreté de la règle) et doit terminer par un nœud déjà visité (pour assurer la fermeture et la connexité).

Pour limiter l'espace de recherche, RuDiK définit également une longueur maximale pour le chemin qui représente le nombre maximal de prédicats dans le corps de la règle.

Enfin, RuDiK ajoute des comparaisons entre les constantes dans le corps de la règle en conservant les comparaisons qui améliorent la qualité de la règle parmi toutes les possibilités construites en considérant toutes les constantes et tous les prédicats binaires suivants : $\{<, >, \leq, \geq, \neq\}$.

Pour mesurer la qualité des règles, RuDiK utilise la mesure du poids détaillée dans la Section 3.3.2

Même si RuDiK découvre des règles comportant des prédicats de comparaison entre constantes, celles-ci doivent être définies dans le graphe de connaissance et associées à deux variables déjà définies dans la règle. L'approche ne permet pas de découvrir une constante de référence comme "*âge(X, a), a* $\geq$ *18* $\rightarrow$ *adulte(X)*", ni de calculer la meilleure valeur de cette constante de référence afin d'améliorer la qualité de la règle.

D'autres approches de type *générer et tester* telles que [8] peuvent être guidées par la sémantique de l'ontologie pour éviter de construire des règles sémantiquement redondantes. Par exemple la règle $pere(x) \wedge parent(x) \rightarrow humain(x)$ est redondante parce que le prédicat $pere$ est un sous-concept de $parent$ (i.e. $pere \sqsubseteq parent$), [8] filtre les règles découvertes en rejetant celles qui sont redondantes. Cependant, l'auteur a montré que l'exploitation des capacités de raisonnement pendant le processus d'apprentissage ne permet pas d'exploiter l'approche sur les grands graphes.

Mesures de qualité

Pour mesurer la qualité des règles, les approches relationnelles peuvent utiliser les mesures de qualité classique de support et de confiance. Le *support* ($supp$) représente le nombre de prédictions correctes de la règle, c.-à-d. le nombre des instanciations correctes de la tête de la règle lorsque le corps de la règle est satisfait. Il est défini par :

$$supp(\overrightarrow{B} \rightarrow h(x,y)) = \#(x,y) : \exists z_1, ..., z_m : \overrightarrow{B} \wedge h(x,y) \tag{3.1}$$

où $z_1, ..., z_m$ sont des variables (ou des constantes) qui instancient $\overrightarrow{B}$.

La confiance ($conf$) est définie par le ratio entre le support de la règle et le nombre de diagnostics différents qui participent à une instanciation du corps de la règle.

$$conf(\overrightarrow{B} \rightarrow h(x,y)) = \frac{supp(\overrightarrow{B} \rightarrow h(x,y))}{\#(x,y) : \exists z_1, ..., z_m : \overrightarrow{B}} \tag{3.2}$$

Le support est difficile à fixer quand certaines conclusions ne concernent que peu d'instances tandis que d'autres

en concernent un grand nombre. Ainsi fixer un seuil minimum de support à 1000 sera plus adapté à des règles concluant sur des personnes exerçant le métier d'informaticien que pour celles exerçant le métier de sénateur. Pour prendre en compte, la faculté d'une règle à générer une proportion significative des exemples instanciant la tête de la règle, la mesure de *head coverage* (hc) a été définie [27]. Plus précisément, le *head coverage* (hc) représente le ratio entre le support et la taille de la tête de la règle (c.-à-d. le nombre des instanciations différentes qui sont présentes dans le graphe de connaissance) :

$$hc(\overrightarrow{B} \rightarrow h(x,y)) = \frac{supp(\overrightarrow{B} \rightarrow h(x,y))}{size(h)} \tag{3.3}$$

D'autres mesures de qualité ont été définies afin de mieux prendre en compte la notion de contre-exemple en monde ouvert. Ainsi, l'hypothèse de Complétude partielle (*Partial Completness Assumption* - PCA) précise que si au moins un individu ou un litéral est défini pour une propriété et un individu donné alors tous les autres individus ou litéraux peuvent être considérés comme des contre-exemples. La PCA-confiance a été définie pour ne tenir compte que de ces contre-exemples dans le calcul de la confiance. De même, la fonction de poids définie par RuDiK [43], permet d'estimer la qualité de la règle à partir des exemples positifs et négatifs qu'elle couvre, en pondérant l'impact des positifs et des négatifs, et en limitant les exemples négatifs à ceux pour lesquels les informations apparaissant dans le corps de règle existent (sans imposer de connexité).

Étant donné un graphe de connaissance GC et le prédicat cible $h(x,y)$ qui apparait dans la tête de la règle. Soit G, l'ensemble de générations qui est constitué de l'ensemble de toutes les paires d'entités (x,y) tel que $h(x,y) \in GC$. Soit V l'ensemble de validation qui se compose de tous les contre-exemples pour le même prédicat cible h. Sachant que $G \cap V = \emptyset$, le poids d'une règle r est défini de la façon suivante :

$$w(r) = \alpha \cdot (1 - \frac{\mid C_r(G) \mid}{\mid G \mid}) + \beta \cdot (\frac{\mid C_r(V) \mid}{\mid U_r(V) \mid}) \tag{3.4}$$

Avec :

— $\alpha, \beta \in [0,1]$ et $\alpha + \beta = 1$ alors $w(r) \in [0,1]$.

— Le C_r d'un ensemble de paires $E = (x_1,y_1),(x_2,y_2),...,(x_n,y_n)$ représente la couverture du corps de la règle r sur E, c.-à-d. $C_r(E)$ est l'ensemble des éléments de E couverts par le corps de r.

— Le $U_r(E)$ dite la couverture déconnectée de l'ensemble E est obtenue en deux étapes :

1. Transformer d'abord le corps de la règle r_{corps} vers r^*_{corps} appelé corps déconnecté. Il est obtenu en ne conservant que les prédicats qui contiennent une variable qui apparaît dans la tête de la règle(i.e x ou y) et en remplaçant les variables qui sont associées à x et y par de nouvelles variables uniques afin de les déconnecter, par exemple : $r_{corps} = pred_1(x,v_0) \wedge pred_2(v_0,y)$ où x et y sont les variables cibles, va être transformé en $r^*_{corps} = pred_1(x,v_i) \wedge pred_2(v_{ii},y)$ (noter que x et y sont déconnectées dans r^*_{corps}).

2. Sélectionner les éléments dans E qui sont couverts par r^*_{corps}.

Le poids représente la qualité d'une règle par rapport à G et V : plus le poids est faible, meilleure est la règle. Une règle parfaite couvre tous les éléments de génération de G et aucun des éléments de validation de V, et obtient alors un poids de 0.

D'autres approches comme [8] utilisent aussi un seuil de fréquence qui représente le taux minimal de co-apparition ou de co-occurence d'un ensemble de concept.

Biais de langage

La découverte de règles dans les graphes de connaissances est une procédure qui peut être très coûteuse quand le graphe est volumineux. Dans le but de limiter l'espace de recherche ou de définir plus précisément l'ensemble des règles qui peuvent être considérées comme un résultat envisageable, les approches de fouille de règles ont utilisé différents biais de langage.

Ces contraintes peuvent limiter les prédicats et les variables qu'il est possible d'ajouter dans le corps ainsi que dans la tête de la règle.

Dans le but d'éviter les règles avec des atomes complètement indépendants, il faut ajouter la contrainte de connexité, c.-à-d. pour qu'une règle soit acceptée, il faut qu'elle soit connexe [27, 43].

Définition 3 (Règle connexe) *Une règle est connexe si chaque atome est connecté transitivement à chaque autre atome de la règle. On dit que deux atomes d'une règle sont connectés s'ils partagent une variable ou une entité.*

Un exemple d'une règle connexe : $Marie(x, y) \wedge Habite(x, z) \rightarrow Habite(y, z)$ où tous les prédicats sont connectés.

Si une approche n'est pas intéressée par la prédiction de règles existentielles qui prédisent l'existence d'un fait sans instancier la variable introduite (p. ex. exemple $person(X) \rightarrow pere(Y, X)$), elle peut utiliser une contrainte imposant que les règles soient fermées [27, 43] :

Définition 4 (Règle fermée) *Une règle est fermée si toutes ses variables sont fermées. Pour qu'une variable dans une règle soit fermée, il faut que celle-ci apparaisse au moins deux fois dans la règle.*

Par exemple toutes les variables dans la règle suivante sont fermées : $pere(x, y) \wedge pere(x, z) \rightarrow frere(y, z)$.

La condition de sûreté peut être utilisée afin de limiter les règles générées à celles pouvant être appliquées et vérifiées [43].

Définition 5 (Règles sûres) *Étant donné une règle $r : B_1, ..., B_n \rightarrow \overrightarrow{H}$ et V l'ensemble des variables, r satisfait la condition de sûreté si toutes les variables apparaissant dans la tête de règle apparaissent également dans le corps de règle, c.-à-d. $V(\overrightarrow{H}) \subseteq \bigcup_{i=1}^{n} V(B_i)$*

Un exemple d'une règle sûre où toutes les variables de la tête x et z apparaissent dans le corps est : $fils(x, y) \wedge fils(z, y) \rightarrow frere(x, z)$. L'utilité de la sûreté d'une règle repose lorsqu'on manipule des règles avec plusieurs prédicats dans la tête où les critères de la fermeture et la connexité ne suffisent pas pour trouver des règles décidables.

Certaines approches permettent de définir les prédicats du GC utilisables dans le corps et la tête de la règle comme AMIE3 [27].

Enfin, pour limiter l'espace de recherche et améliorer l'interprétabilité des règles, il est aussi possible de définir également une longueur maximale de la règle qui correspond au nombre total de prédicats apparaissant dans le corps et la tête de la règle, comme cela est fait dans [27] et [43].

L'utilisation de seuils pour une ou plusieurs mesures de qualité peut également être vue comme un biais de langage. Par exemple, la mesure de *head coverage* hc d'une règle (défini dans la section 3.3.2) peut servir à limiter l'espace de recherche et arrêter de développer une règle quand le seuil n'est pas atteint.

3.4 Discussion

Les approches d'induction de règles présentées dans ce chapitre découvrent des règles dont la complexité peut varier et ces approches suivent des stratégies différentes pour explorer l'espace de recherche. Le tableau 3.1 synthétise les différentes caractéristiques des approches étudiées en présentant le type de règles logique pouvant être apprises ainsi que la possibilité d'utiliser la négation, des constantes, de comparer deux valeurs numériques décrivant des variables, ou comparer la valeur associée à une variable avec une valeur de référence calculée. Le tableau présente également la stratégie et les mesures de qualité utilisées, les contraintes exprimables grâce au biais de langage, et l'utilisation de la sémantique de l'ontologie.

Toutes les approches présentées découvrent des clauses de Horn, mais avec des caractéristiques différentes. La composition des règles diffère en effet d'une approche à l'autre. AMIE3 [27], RuDiK [43] et TILDE [2] permettent l'utilisation des constantes dans les règles. L'utilisation de la négation est considérée par RuDiK pour lui permettre de trouver des règles de validation d'un graphe de connaissances. Enfin, RuDiK et TILDE autorisent l'utilisation des prédicats de comparaisons dans les règles, mais seul TILDE calcule des valeurs de référence qui servent à seuiller des variables numériques durant la construction de l'arbre de décision. Aucune des approches de type *Générer et tester* n'a envisagé cette possibilité jusqu'à présent. Les approches de fouille de règles dans les graphes de connaissance sont très efficaces pour découvrir des règles comportant peu d'atomes en prémisse (c.-à-d. par défaut, AMIE3 ne considère que trois atomes). Quand on cherche à découvrir des règles comportant potentiellement un grand nombre d'atomes, le biais de langage peut guider l'exploration de l'espace de recherche en définissant les chemins et les relations à explorer. Ce guide peut être réalisé pour définir certaines caractéristiques de la règle (c.-à-d. connexité et fermeture) ou pour définir les prédicats à explorer pour construire une règle significative.

Ces biais peuvent se limiter à la définition de la conclusion et des prédicats à considérer comme dans [27], ou permettre la déclaration de modes plus complexes comme [2] qui permet à la fois de définir les relations, leur nombre d'apparitions, ou encore le type de leurs arguments.

Nous avons aussi montré que ces approches utilisent des mesures de qualité différentes pour évaluer les règles découvertes. Certaines sont directement inspirées des approches de fouille de règles d'association classiques comme le support et la confiance. D'autres mesures permettent de s'adapter à la distribution des données entre les différentes conclusions considérées comme le *head coverage*, tandis que d'autres mesures combinent la couverture et l'exactitude comme le poids, pour capturer la qualité d'une règle par rapport à sa couverture des exemples positifs et négatifs.

[8] exploite la sémantique de l'ontologie, et plus précisément la subsumption, pour éliminer les règles sémantiquement redondantes après leur génération. A notre connaissance, aucune approche de fouille de règles dans les graphes de connaissances n'utilise la sémantique de l'ontologie durant le processus de découverte.

Ces approches suivent des stratégies d'induction différentes pour répondre à différents besoins. Certaines utilisent une stratégie de type Divide-and-Conquer pour découvrir les règles les plus précises pour chaque classe d'exemples comme [2]. D'autres veulent couvrir plus de possibilités et de combinaisons de prédicats en utilisant des stratégies de type Generate-and-Test, comme [27, 8].

Chacune de ces approches considère différemment les exemples non classés (qui ne sont pas déclarés ni comme positifs ni comme négatifs), AMIE3 [27] par exemple surmonte la difficulté du manque de contre-exemples en utilisant les techniques de PCA qui lui permettre de deviner des contre-exemples, ces contre-exemples devinés vont être considérés lors le calcule de la confiance d'une règle en utilisant la PCA-confiance. Tandis que les autres approches [43, 8, 2] suivent l'hypothèse du monde ouvert et ne prennent aucune supposition sur la nature des exemples inconnus.

Dans l'état de l'art, il n'existe pas une approche qui permet :

1. d'utiliser les heuristiques dédiées aux relations partie-tout pour utiliser le contexte dans lequel les composants ont participé, et pour montrer l'influence du contexte sur la précision des règles. Nous voulons exprimer le contexte défini par l'expert d'amiante à CSTB dans un biais de langage. Ce contexte est essentiellement composé par des relations partie-tout entre les composants du bâtiment.

2. d'intégrer des contraintes calculées sur des valeurs numériques qui représentent des informations temporelles. Ces contraintes sont calculées durant l'exploration des règles et ses valeurs doivent maximisent la précision des règles. Les valeurs des contraintes calculées participent dans des comparaisons avec les variables numériques dans le graphe de connaissances comme les données temporelles.

3. de prédire des règles en utilisant la sémantique d'ontologie afin de détecter la hiérarchie des contextes et distinguer entre les contextes et les sous-contextes. L'objectif principal de l'utilisation de la sémantique est

de limiter l'espace de recherche. C.-à-d. limiter l'exploration *Top-down* des règles.

Dans cette problématique, nous cherchons des règles contextuelles et sémantiques avec des contraintes calculées. Aucune des approches existantes aujourd'hui n'est capable de fournir ce type de règle. En plus, l'objectif dans cette problématique est de trouver toutes les règles (afin de montrer à l'expert toutes les explications possibles pour chaque composant) contrairement aux approches de type Divide-and-Conquer comme TILDE. Nous avons alors proposé une approche de fouille de règle de type Generate-and-Test qui satisfait les trois points mentionnés afin de résoudre notre problématique.

L'autre approche que nous avons proposée est une approche de classification basée sur les ressources externes contenant des données temporelles, incomplètes et contradictoires. Ils existent plusieurs approches de liage de données, cependant, comme nous ne disposons que des types (classes) des entités et de l'année de construction qui ne suffisent pas d'identifier les composants pour les lier, aucune de ces approches de liage de données se base sur ces deux types d'information pour déduire une classification d'un individu. Cette approche utilise les informations disponibles sur les données (le type et l'année) pour calculer d'abord une probabilité qu'un individu appartienne à une classe, puis elle apprend un seuil de probabilité pour classer l'individu.

3.5 Conclusion

Dans ce chapitre nous avons tout d'abord montré que les différentes méthodologies se différenciaient en fonction de leur automatisation, de leur capacité à prendre en compte la réutilisation de ressources ontologiques ou non structurées, et les aspects collaboratifs. Dans le domaine du bâtiment, il n'existe pas d'ontologies qui décrivent les bâtiments et les éléments qui le composent en respectant les éléments décrits par la norme utilisée par le CSTB. L'objectif est de modéliser les bâtiments et les diagnostics avec l'aide d'un expert, en suivant cette norme et en réutilisant des ressources ontologiques pour modéliser les informations temporelles et des ressources externes non-structurées qui décrivent les produits amiantés. La méthodologie que nous avons appliquée s'inspire donc plutôt d'une méthodologie telle que NeOn.

Nous avons ensuite présenté les stratégies d'induction de règles qui cherchent à découvrir des hypothèses de type clause de Horn en utilisant un ensemble de connaissances et un ensemble d'exemples positifs et négatifs. Nous avons montré que les approches existantes varient en fonction des stratégies de parcours de l'espace de recherche, des biais de langage pouvant être déclarés, des hypothèses de complétude posées sur les données et des mesures de qualité utilisées. Dans le cadre de notre projet, nous voulons explorer deux types de stratégies. Dans la première stratégie, l'objectif est d'utiliser des ressources externes décrivant les produits commercialisés et la présence d'amiante dans ces produits en fonction des années pour calculer des probabilités de présence d'amiante par classe de produit et par année afin de les utiliser dans des règles déclarées qui se base sur ces probabilités, sur les informations temporelles associées à un produit et sur un seuil pour prédire la présence d'amiante. Ce

Approches d'induction	AMIE3 [27]	RuDiK [43]	TILDE [2]	D'Amato et al. [8]
Type de règle	Clauses de Horn	Clauses de Horn	Arbre de décision relationnel (séquence de SI ALORS SINON)	Clauses de Horn
Négation	NON	OUI	NON	NON
Constantes	Numériques Littéraux	Numériques Littéraux	Numériques Littéraux	Numériques Littéraux
Comparaison	NON	OUI	OUI	NON
Comparaison avec des valeurs de référence	NON	NON	OUI	NON
Sémantique de l'ontologie	NON	NON	NON	OUI (filtrer les redondances)
Hypothèse de l'espace de recherche	PCA	OWA	OWA	OWA
Type de stratégie	Generate-and-Test	Generate-and-Test	Divide-and-Conquer	Generate-and-Test
Mesures de qualité	Support *Head coverage* Confiance	Poids	/	Support *Head coverage* Confiance Fréquence
Biais de langage	Ensemble de relations Fermeture Connexité Longueur maximale	Fermeture Connexité Sûreté Longueur maximale	Déclaratif : Utiliser les modes pour préciser les relations à ajouter, leur nombre d'apparitions et le type de leurs arguments (variables ou constantes)	Ensemble de relations Fermeture Connexité Sûreté Redondance Longueur maximale

TABLE 3.1 – Comparaison entre les approches d'induction de règles dans les graphes de connaissances

seuil peut être appris sur un ensemble peu volumineux de diagnostics. Dans une deuxième stratégie, l'objectif est d'exploiter l'ontologie peuplée par les diagnostics pour induire des règles FOL concluant sur la présence d'amiante en utilisant le contexte dans lequel il est utilisé et les informations temporelles. Nous souhaitons conserver toutes les règles dont la qualité a été validée afin de montrer à l'expert amiante au CSTB toutes les explications amenant à une prédiction donnée. Dans ce chapitre, nous avons montré que les approches de type Generate-and-Test existantes permettent d'atteindre cet objectif, mais les langages de règles considérés ne permettent pas de calculer des constantes de référence afin de les utiliser dans des comparaisons avec des variables numériques de la règle. Enfin, aucune approche de ce type ne permet d'utiliser la sémantique de l'ontologie pour limiter l'espace de recherche, ou de définir un biais de langage utilisant les propriétés partie-tout pour explorer les relations entre les différents composants d'un bâtiment tout en limitant la taille et la portée des contextes considérés. Même s'il existe des approches qui satisfont quelques points essentiels de notre problème, il n'existe pas d'approche adaptée au problème de prédiction de présence d'amiante qui permette de fournir toutes les fonctionnalités attendues.

Chapitre 4

Construction d'une ontologie de l'Amiante

4.1 Introduction

La première étape de cette thèse avait pour objectif de construire une ontologie appelée ASBESTOS permettant au CSTB de représenter les bâtiments, les diagnostics amiante effectués sur certains éléments de bâtiments.

L'ontologie construite comporte différents modules afin de faciliter la réutilisation partielle de l'ontologie par les experts du domaine et leur permettre d'utiliser ou non les connaissances éventuellement issues de ressources externes au CSTB, ou les connaissances générées par les approches de prédiction que nous avons développées.

Comme il n'existe pas d'ontologie permettant de décrire les bâtiments qui soit adaptée à ce domaine, l'objectif est d'utiliser tout d'abord les documents disponibles au sein du CSTB pour construire cette ontologie en collaboration avec les experts du CSTB. Les documents dont dispose le CSTB sont de deux types : (1) des projets type homologués (documents textuels retranscrits manuellement à partir de documents PDF) et (2) des diagnostics amiante effectués sur des éléments de bâtiments (documents excel). Dans les deux cas, les données d'intérêt sont présentées sous forme de données tabulaires dont la structure est régulière et utilise un vocabulaire similaire à celui défini par la norme NF X46-020 [1] datant de 2017. La structure tabulaire régulière permet d'éviter de devoir faire appel à des outils de découverte de relations. Enfin, les termes décrivant les instances de concepts ne sont pas ambigus et le vocabulaire utilisé est relativement homogène.

Aussi, l'objectif est tout d'abord de définir le haut niveau des différents modules de l'ontologie en collaboration avec l'expert du CSTB et de proposer un processus d'extraction semi-automatique des informations présentes dans les documents du CSTB afin d'enrichir cette ontologie. Pour cela, nous avons défini un processus d'extraction qui permet d'identifier les termes décrivant les concepts puis les relations de subsumption qui les relient en nous basant sur les textes de projets types homologués ou les diagnostics. Ce processus est adapté aux tableaux et aux termes

1. LanormeNFX46-020:https://www.boutique.afnor.org/fr-fr/norme/nf-x46020/reperage-amiante-reperage-des-materiaux-et-produits-contenant-de-lamiante-d/fa186482/1669

utilisés dans ce contexte et aux variations lexicales que l'on peut rencontrer.

Dans ce chapitre, nous décrivons tout d'abord un bref état des lieux sur l'utilisation de l'amiante en France et son repérage dans les bâtiments, puis les ressources dont le CSTB dispose. Nous présentons ensuite l'Ontologie ASBESTOS, ainsi que les différents outils développés pour l'enrichissement semi-automatique de cette ontologie avec de nouvelles classes et son peuplement avec les descriptions de bâtiments issus des diagnostics ou des projets homologués.

4.2 L'Amiante et son repérage dans les bâtiments

Qu'est-ce que l'amiante

L'amiante est un terme générique qui désigne différents silicates fibreux qui existent dans la nature et qui sont de type serpentines ou amphiboles. Ces matériaux ont de nombreuses propriétés : résistance au feu, coût peu élevé, résistance aux agressions chimiques et physiques. Aussi, l'amiante a été utilisé pour l'isolation thermique, la fabrication de joints, de colles, de câbles électriques, de canalisations, ou encore de plaquettes de frein de voiture. Des fibres d'amiante ont également été incorporées aux bitumes ou aux ciments.

Quelques Données santé

L'amiante constitue un problème majeur de santé publique. En effet, ce matériau qui a des qualités multiples est particulièrement toxique.

Le contact direct avec des produits contenant de l'amiante ou l'inhalation des particules flottant dans l'air de ces produits peuvent provoquer des maladies très graves comme le cancer du poumon, mésothéliome, les plaques pleurales et pleins d'autres cancers qui affectent le larynx, les ovaires, le colon, le rectum ou encore l'estomac.

L'amiante a été massivement utilisé en France jusque 1997, année où il a été interdit. Le nombre de cancers que l'amiante a provoqué ne cesse d'augmenter. Ainsi, d'après un rapport du Haut conseil de la santé publique de 2014 [9], au total, sur la période 1955-2009, le nombre de décès attribuable à une exposition à l'amiante serait compris entre 61300 et 118400 (exposition professionnelle uniquement pour le cancer du poumon ; tout type d'exposition pour le mésothéliome). De plus, il faut s'attendre entre 2009 et 2050 à un nombre de décès par cancer du poumon dus à l'amiante de l'ordre de 50 à 75000, auxquels s'ajoutent 18 à 25000 décès dus au mésothéliome, sans même compter les autres types de cancers pour lesquels la responsabilité de l'amiante a été confirmée.

Bien qu'interdit depuis 1997, ce matériau est encore présent dans de nombreux bâtiments.

Repérage de l'amiante dans les bâtiments et Norme AFNOR NF X46-020

Le repérage est une opération effectuée par un opérateur certifié de repérage. Le repérage vise à rechercher, identifier et localiser dans les bâtiments, les matériaux et produits contenant de l'amiante. Le repérage comprend :

— la recherche de matériaux ou produits figurant sur des listes réglementaires ;

— l'identification de la présence ou non d'amiante dans les matériaux précédemment trouvés ;

Le décret du 9 mai 2017 R4412-97 [2] décrit l'obligation de la recherche d'amiante faite au donneur d'ordre (DO), au maître d'ouvrage (MOA) ainsi qu'au propriétaire d'immeubles par nature ou par destination, d'équipements, de matériels ou d'articles qui décide d'une opération comportant des risques d'exposition des travailleurs à l'amiante.

L'AFNOR [3] a défini la norme NF X46-020 [4] afin de guider les missions de repérage d'amiante dans les bâtiments. Cette norme fournit des méthodologies de repérage qui respectent la réglementation française (c.-à-d. le code de la santé publique et le code de la construction et de l'habitation).

La norme et le décret décrivent également le vocabulaire à utiliser pour décrire les bâtiments, leurs parties (c.-à-d. les structures et les localisations) et les classes de produits lors des missions de repérage. Ce vocabulaire normé a été utilisé dans le projet ORIGAMI pour créer une table de référence définissant la structuration d'un bâtiment, et cette référence est respectée par les documents du CSTB comme les diagnostics amiante et les projets types homologués.

Ce vocabulaire indique qu'un bâtiment est composé de plusieurs structures et liste les structures possibles (par exemple : parois verticales intérieures, couvertures, toitures, terrasses, .etc). Chaque structure contient plusieurs localisations appelées aussi composants. La liste des localisations comporte des termes comme : fenêtres, lanterneaux, verrières, façades légères, murs, .etc. Une localisation est composée par plusieurs produits nommés aussi partis du composant. La liste des produits comporte par exemple les termes suivants : revêtements, joints, colle, flocages, enduits projetés, revêtement bitumineux, mastics, .etc.

4.3 Les initiatives dans le domaine du bâtiment

La démarche BIM [22] (Building Information Model/Modeling/Management) est une nouvelle façon de décrire le bâtiment, qui permet de regrouper au sein d'un ensemble de fichiers numériques les informations techniques de l'ouvrage. Elle permet l'élaboration de maquettes numériques, qui intègrent les informations géométriques des ouvrages. L'objectif est également d'intégrer des données sémantiques permettant de qualifier les composants de l'ouvrage. Le BIM définit ainsi le processus de stockage, de génération, de gestion, d'échange et de partage d'information du bâtiment par des acteurs multi-métiers (AEC :Architecture, Engineering, Construction). Ainsi avec

2. Le journal officiel du décret R4412-97 : `https://www.legifrance.gouv.fr/eli/jo/2019/7/18/0165`
3. L'AFNOR (`https://www.afnor.org/`) est une association qui constitue un groupe international au service de l'intérêt général et du développement économique.
4. La norme NF X46-020 : https://www.boutique.afnor.org/fr-fr/norme/nf-x46020/reperage-amiante-reperage-des-materiaux-et-produits-contenant-de-lamiante-d/fa186482/1669

l'essor de la maquette numérique, le BIM est devenu une pièce maîtresse pour l'aménagement du territoire et la conception du bâtiment.

Les fichiers numériques créés par le projet BIM décrivent des propriétés sur l'architecture des bâtiments comme les surfaces des espaces dans les bâtiments, la hauteur des murs, la longueur des tuyaux, ou encore les propriétés 3D de la maquette numérique du plan du bâtiment. L'exemple [5] suivant représente la propriété couleur d'un mur :

— Objet : MUR

— Propriété : couleur

— Valeur de la propriété : rouge

Plusieurs initiatives dans le monde ont vu le jour pour faciliter l'échange entre les différents acteurs du bâtiment tel que ifcOWL [6] qui se base sur les documents de BIM pour définir une première version d'ontologie du bâtiment (draft 2019).

Cependant, les propriétés décrites dans les documents BIM et dans l'ontologie ifcOWL ne sont pas exploitables pour représenter les données des diagnostics amiante et pour la problématique de prédiction de présence d'amiante dans les composants du bâtiment. En effet, ceux-ci décrivent les propriétés de la maquette et le plan du bâtiment, mais ne décrivent ni les relations entre les composants dans le bâtiment ni les produits utilisés dans ces composants.

Par exemple le concept "ifc window type" décrit le type de la fenêtre utilisée dans un bâtiment, mais ne décrit pas la composition de cette fenêtre, c.-à-d. les produits qui compose cette fenêtre comme le bois, la peinture, etc. De plus, à notre connaissance, aucune initiative autre que ifcOWL n'a abouti à la mise à disposition d'une ontologie.

Les travaux actuels ne s'intéressent donc qu'aux caractéristiques architecturales du bâtiment pour les numériser et les partager. Ils ne permettent pas de représenter la composition détaillée du bâtiment qui est jugée nécessaire par l'expert afin de prédire la présence d'amiante.

4.4 Ressources disponibles

Durant la construction de l'ontologie, nous avons utilisé certaines des ressources documentaires dont dispose le CSTB ainsi que la ressource ontologique OWL-Time afin de représenter les données temporelles.

4.4.1 Documents du CSTB

Les ressources du CSTB sont des ressources non-structurées qui comportent des données présentées de manière tabulaire. Il s'agit d'un ensemble de documents archivés par le CSTB qui décrivent les bâtiments construits

5. DEVELOPPEMENT D UN DICTIONNAIRE DE PROPRIETES/OUVRAGES ET D UNE BIBLIOTHEQUE DE MODELES D OBJETS GENERIQUES BIM : http://docplayer.fr/73536652-Developpement-d-un-dictionnaire-de-proprietes-ouvrages-et-d-une-bibliotheque-de-modeles-d-objets-generiques-bim.html

6. L'ontologie ifcOWL : https://standards.buildingsmart.org/IFC/DEV/IFC4/ADD2_TC1/OWL/index.html

en France. Ils sont de deux types :

— **Diagnostics amiante** : appelé également Rapport Avant Travaux (RAT), décrivent les résultats des diagnostics effectués pour détecter la présence d'amiante dans des parties de bâtiments (c.-à-d. résultats d'analyse sur des prélèvements). L'ensemble des données des RAT ont été agrégées par le CSTB dans un document Excel global appelé Diagnostics amiante. Le fichier Excel décrit les résultats de tests. Il contient un ensemble d'informations décrivant le bâtiment : son nom, son adresse, la région, son type (p. ex. école, maison individuelle, immeuble collectif, hôpital), et l'année de construction du bâtiment. Il contient également les structures qui composent le bâtiment (p. ex. ouverture extérieure, balcon). Pour chaque structure, le document décrit l'ensemble de ses localisations (p. ex., porte, fenêtre), et pour chaque localisation il mentionne les types de produits utilisés (p. ex. enduit, colle, etc.). De plus, le document décrit les résultats des tests sur les produits. La figure 4.1 montre un extrait de diagnostics. La première ligne représente les entêtes du tableau. Chaque ligne présente un produit et un ensemble d'information en lien avec ce produit. Elle commence par l'identifiant du bâtiment (id_bat). Nous trouvons ensuite l'identifiant du produit (id_p_comp), celui de la structure (id_struct) et celui de la localisation (id_comp). Les deux colonnes suivantes contiennent le type de la pièce (c.-à-d. séjour, salle de bain) et du bâtiment (immeuble collectif ou individuel). Ensuite, on trouve le code postale, la commune et l'année de construction du bâtiment. La dernière colonne contient le résultat du diagnostic amiante pour le produit représenté dans la ligne : "0" pour l'absence d'amiante et "1" pour sa présence. Ainsi, la deuxième ligne décrit un bâtiment dont l'identifiant est "111002" qui contient dans une structure de type "Parois verticales intérieures", une localisation "Cloisons sèches", et cette localisation comporte un produit de type "Bandes calicot" qui est amianté. Il est aussi mentionné que ce bâtiment est localisé dans la commune "PALAISEAU" associée au code postal "91120" et il est construit en 1986.

— **Projet type homologué** : document qui contient un ensemble d'informations décrivant un type de bâtiment qui a pu être construit en plusieurs exemplaires et la liste des bâtiments construits en suivant ce projet type homologué avec leur année de construction. Un projet type décrit comme dans un RAT la liste de ses structures, de ses localisations, ainsi que les types de produits utilisés. Les figures 4.2 et 4.3 montrent un exemple de projet type homologué et sa retranscription manuelle dans un fichier .csv. L'objectif étant pour le CSTB de numériser automatiquement ces documents par la suite, pour les exploiter. La partie haute de la Figure 4.2 montre un tableau qui contient les informations du bâtiment, information retranscrite dans la première ligne de la Figure 4.3 : le type de logement "F3" , la région "ÎLE-DE-FRANCE", le numéro de département "75", le numéro d'enregistrement "6", le nom et l'adresse du bénéficiaire de l'homologation. La partie "Résumé du descriptif" de la Figure 4.2 comporte un tableau de deux colonnes, la première contient le nom de la structure et la deuxième contient une description textuelle de l'ensemble des localisations et de classes de produit utilisées dans la structure. Cette structure a été retranscrite vers une structure

id_bat	id_p_comp	id_struct	id_comp	type_piece	type_bat	cd_postal	commune	annee_construction	presence_amiante
111002	Bandes calicot	Parois verticales intérieures	Cloisons sèches (asse	Dégagement	Immeuble collectif	91120	PALAISEAU	1986	1
111002	Bandes calicot	Parois verticales intérieures	Cloisons sèches (asse	Placard	Immeuble collectif	91120	PALAISEAU	1986	1
111002	Enduits à base de plâtre ou de ciment projetés, lissés (	Parois verticales intérieures	Murs et cloisons maç	Entrée	Immeuble collectif	91120	PALAISEAU	1986	1
111002	Colles et joints de carrelage ou de faïence, ragréage, p	Parois verticales intérieures	Revêtements de murs	Cuisine	Immeuble collectif	91120	PALAISEAU	1986	1
111002	Enduits de ragréage, débullage, lissage	Parois verticales intérieures	Murs et cloisons maç	Entrée	Immeuble collectif	91120	PALAISEAU	1986	1
111002	Bandes calicot	Parois verticales intérieures	Cloisons sèches (asse	Cuisine	Immeuble collectif	91120	PALAISEAU	1986	1
111002	Bandes calicot	Parois verticales intérieures	Cloisons sèches (asse	Chambre 3	Immeuble collectif	91120	PALAISEAU	1986	1
111002	Enduits à base de plâtre ou de ciment projetés, lissés (	Parois verticales intérieures	Murs et cloisons maç	Chambre 1	Immeuble collectif	91120	PALAISEAU	1986	1
111002	Bandes calicot	Parois verticales intérieures	Cloisons sèches (asse	Cellier	Immeuble collectif	91120	PALAISEAU	1986	1
111002	Bandes calicot	Parois verticales intérieures	Cloisons sèches (asse	Entrée	Immeuble collectif	91120	PALAISEAU	1986	1
111002	Bandes calicot	Parois verticales intérieures	Cloisons sèches (asse	Séjour	Immeuble collectif	91120	PALAISEAU	1986	1
111002	Enduits à base de plâtre ou de ciment projetés, lissés (	Parois verticales intérieures	Murs et cloisons maç	Placard	Immeuble collectif	91120	PALAISEAU	1986	1
111002	Bandes calicot	Parois verticales intérieures	Cloisons sèches (asse	Placard	Immeuble collectif	91120	PALAISEAU	1986	1
111002	Bandes calicot	Parois verticales intérieures	Cloisons sèches (asse	Placard	Immeuble collectif	91120	PALAISEAU	1986	1
111002	Enduits de ragréage, débullage, lissage	Parois verticales intérieures	Murs et cloisons maç	Chambre 1	Immeuble collectif	91120	PALAISEAU	1986	1
111002	Enduits de ragréage, débullage, lissage	Parois verticales intérieures	Murs et cloisons maç	Séjour	Immeuble collectif	91120	PALAISEAU	1986	1
111002	Bandes calicot	Parois verticales intérieures	Cloisons sèches (asse	Chambre 2	Immeuble collectif	91120	PALAISEAU	1986	1
111002	Bandes calicot	Parois verticales intérieures	Cloisons sèches (asse	WC	Immeuble collectif	91120	PALAISEAU	1986	1

FIGURE 4.1 – Extrait des diagnostics amiante

tabulaire similaire. Ce bâtiment contient 13 structures. Lorsqu' une structure contient plusieurs localisations, les localisations et les produits sont séparés par un "-" (p. ex. la structure I contient 4 localisations séparées par des "-"). Dans le cas où il y a plusieurs localisations qui partagent la même composition de produits, elles sont mentionnées ensembles et séparées des produits par des " :" comme dans l'exemple de la structure C. La localisation peut ne pas être mentionnée comme le cas de la structure B. Certaines structures sont inexistantes et notées Néant, comme c'est le cas pour les structures F et J.

Ces deux type de documents respectent les éléments de description des bâtiments définis dans la Norme NF X46-020 [7] définie par l'AFNOR.

L'ontologie ASBESTOS est construite à partir de ces ressources documentaires, en utilisant celles qui couvrent la période de 1943 à 1997 (c.-à-d. date d'interdiction de l'amiante).

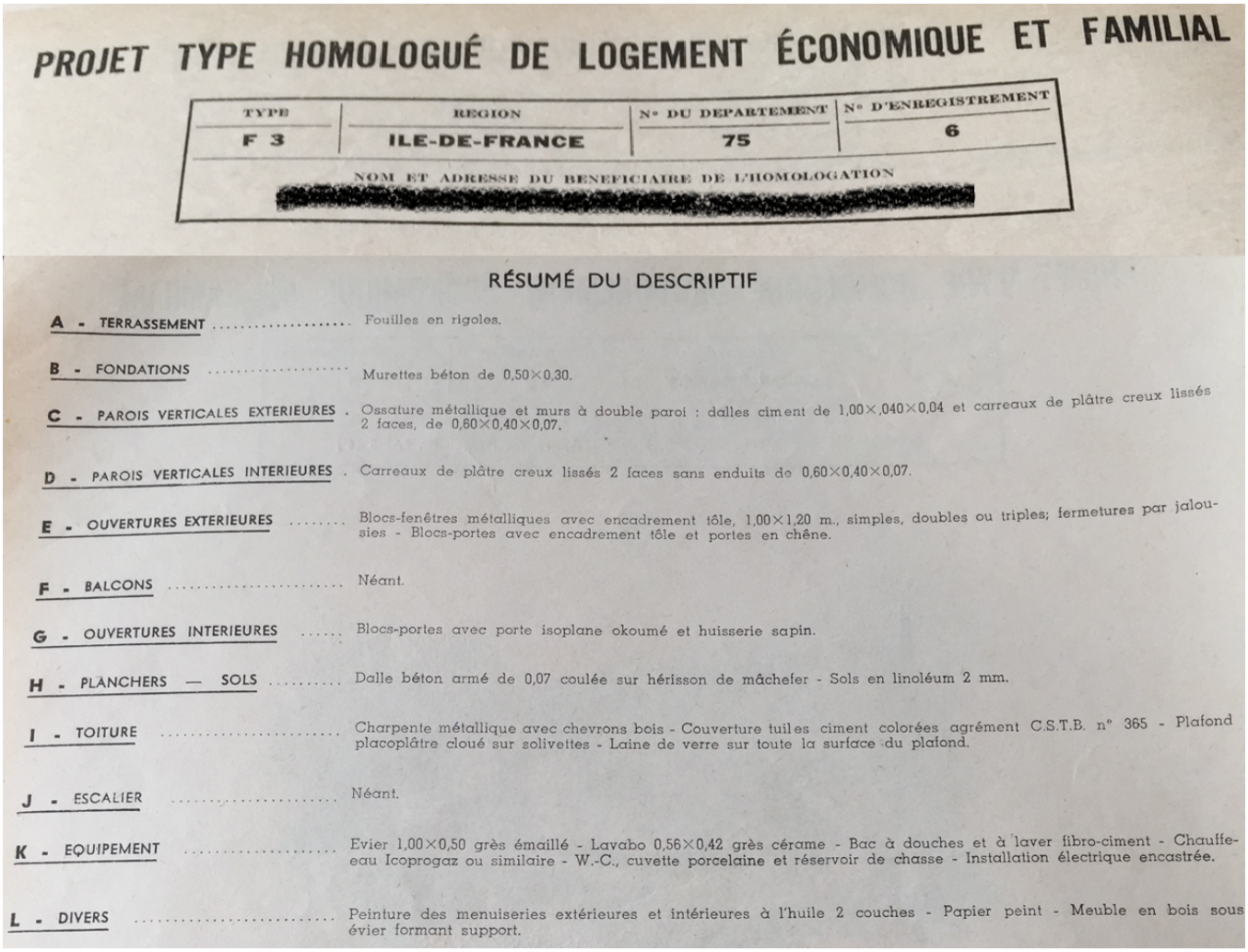
4.4.2 Ressources ontologiques

Parmi les données représentées dans les ontologies, certaines données sont temporelles. Ainsi, dans des applications comme celle du CSTB, il est nécessaire de représenter des dates de construction de bâtiments, mais également des intervalles de temps dans lesquels des produits ont été amiantés et de pouvoir éventuellement raisonner sur ces intervalles.

La Time Ontology [8] (ou OWL-Time) a été définie pour permettre de représenter des données temporelles complexes. Elle permet en effet la représentation des instants et des intervalles de temps.

Cette ontologie du temps définit une entité temporelle [9] (Figure 4.4) comme une entité ayant différentes propriétés telles que : instant de début, de fin, une durée, etc. Un instant est une entité temporelle caractérisée par des propriétés telles que : année, mois, jour, ou heure, dont le début et la fin représentent le même instant et pour lequel la durée est nulle. Un intervalle de temps est une entité temporelle qui est décrite en particulier par un instant

7. La norm NF X46-020 : `https://www.boutique.afnor.org/fr-fr/norme/nf-x46020/reperage-amiante-reperage-des-materiaux-et-produits-contenant-de-lamiante-d/fa186482/1669`
8. Ontologie du temps dans OWL : `https://www.w3.org/TR/owl-time/`
9. Les relations temporelles topologiques dans OWL-Time : `https://www.w3.org/TR/owl-time/#topology`



PROJET TYPE HOMOLOGUÉ DE LOGEMENT ÉCONOMIQUE ET FAMILIAL

TYPE	REGION	N° DU DEPARTEMENT	N° D'ENREGISTREMENT
F 3	ILE-DE-FRANCE	75	6

NOM ET ADRESSE DU BENEFICIAIRE DE L'HOMOLOGATION

[redacted]

RÉSUMÉ DU DESCRIPTIF

A - TERRASSEMENT	Fouilles en rigoles.
B - FONDATIONS	Murettes béton de 0,50×0,30.
C - PAROIS VERTICALES EXTERIEURES	Ossature métallique et murs à double paroi : dalles ciment de 1,00×,040×0,04 et carreaux de plâtre creux lissés 2 faces, de 0,60×0,40×0,07.
D - PAROIS VERTICALES INTERIEURES	Carreaux de plâtre creux lissés 2 faces sans enduits de 0,60×0,40×0,07.
E - OUVERTURES EXTERIEURES	Blocs-fenêtres métalliques avec encadrement tôle, 1,00×1,20 m., simples, doubles ou triples; fermetures par jalousies - Blocs-portes avec encadrement tôle et portes en chêne.
F - BALCONS	Néant.
G - OUVERTURES INTERIEURES	Blocs-portes avec porte isoplane okoumé et huisserie sapin.
H - PLANCHERS — SOLS	Dalle béton armé de 0,07 coulée sur hérisson de mâchefer - Sols en linoléum 2 mm.
I - TOITURE	Charpente métallique avec chevrons bois - Couverture tuiles ciment colorées agrément C.S.T.B. n° 365 - Plafond placoplâtre cloué sur solivettes - Laine de verre sur toute la surface du plafond.
J - ESCALIER	Néant.
K - EQUIPEMENT	Evier 1,00×0,50 grès émaillé - Lavabo 0,56×0,42 grès cérame - Bac à douches et à laver fibro-ciment - Chauffe-eau Icoprogaz ou similaire - W.-C., cuvette porcelaine et réservoir de chasse - Installation électrique encastrée.
L - DIVERS	Peinture des menuiseries extérieures et intérieures à l'huile 2 couches - Papier peint - Meuble en bois sous évier formant support.

FIGURE 4.2 – Extrait d'un projet type homologué

F3	ILE-DE-France	75	6	[redacted]	[redacted]	1970
TERRASSEMENT	Fouilles en rigoles.					
FONDATIONS	Murettes béton de 0,50*0,30.					
PAROIS VERTICALES EXTERIEURES	Ossature métallique et murs à double paroi : dalles ciment de 1,00*,040*0,04 et carreaux de plâtre creux lissés 2 faces, de 0,60*0,40*0,07.					
PAROIS VERTICALES INTERIEURES	Murs avec carreaux de plâtre creux lissés 2 faces sans enduits de 0,60*0,40*0,07.					
OUVERTURES EXTERIEURES	Blocs fenêtres métalliques avec encadrement tôle, 1,00*1,20 m., simples, doubles ou triples; fermetures par jalousies - Blocs ports avec encadrement tôle et portes en chêne.					
BALCONS	Néant.					
OUVERTURES INTERIEURES	Blocs portes avec porte isoplane okoumé et huisserie sapin.					
PLANCHERS-SOLS	Dalle béton armé de 0,07 coulée sur hérisson de mâchefer - Sols en linoléum 2 mm.					
TOITURE	Charpente métallique avec chevrons bois - Couverture tuiles ciment colorées agrément C.S.T.B. n° 365 - Plafond placoplâtre cloué sur solivettes - Laine de verre sur toute la surface de plafond.					
ESCALIER	Néant.					
EQUIPEMENT	Evier 1,00*0,50 grès émaillé - Lavabo 0,56*0,42 grès cérame - Bac à douches et à laver fibro-ciment - Chauffe-eau Icoprogaz ou similaire - WC, cuvette porcelaine et réservoir de chasse - Installation électrique encastrée.					
DIVERS	Peinture des menuiseries extérieures et intérieures à l'huile 2 couches - Papier peint - Meuble en bois sous évier formant support.					

FIGURE 4.3 – Retranscription du projet type homologué présenté dans la Figure 4.2

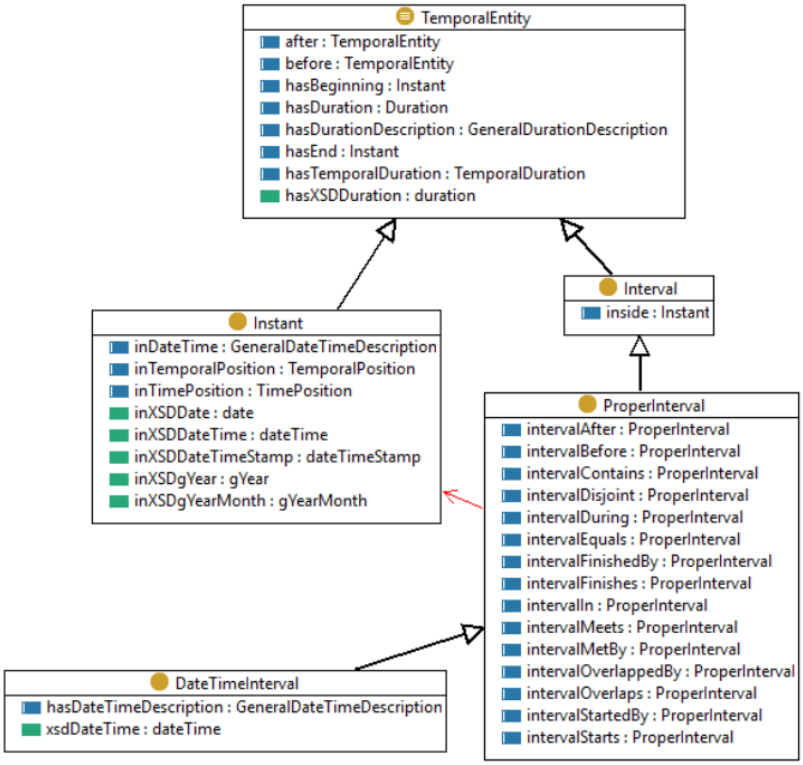


FIGURE 4.4 – Modèle de base des entités temporelles défini par OWL-Time

de début, de fin et une durée et par un ensemble de relations possibles avec d'autres intervalles de temps (e.g. *intervalMeets*, *intervalDisjoint*, *intervalOverlappedBy*).

Quand les données temporelles sont définies en suivant ce vocabulaire, qui respecte la logique temporelle définie par Allen, il est possible de faire des raisonnements en utilisant les intervalles et les instants. Par exemple, on peut savoir si un évènement daté se situe dans un intervalle de temps spécifié.

4.5 Les méthodologies de construction d'ontologies

La construction d'ontologie est une tâche difficile et très coûteuse en termes de temps de réalisation. Une ontologie peut être créée (1) à partir de zéro, (2) à partir d'ontologies existantes, (3) à partir d'un corpus de sources d'informations non structurées, ou (4) en combinant ces deux dernières approches.

Il existe deux méthodes [12] de construction d'une ontologie : la construction manuelle et la construction automatique basée sur des méthodes de fouille de données. La construction manuelle d'ontologie suppose des interactions avec un ou plusieurs experts du domaine et propose éventuellement des méthodologies collaboratives. Cependant, un processus manuel peut être chronophage et peut poser des problèmes de mise à jour. D'autre part, les méthodes de construction automatiques ne fonctionnent bien que pour des ontologies peu expressives dans des domaines très limités. Aussi, de nombreux systèmes sont semi-automatiques. Ces systèmes peuvent exploiter

Méthodologies de construction d'ontologie	METHONTOLOGY	On-To-Knowledge	DILIGENT	NeOn
Réutilisation de ressources ontologique	NON	OUI	OUI	OUI
Guide de réutilisation de ressources ontologique	NON	NON	NON	OUI
Ingénierie des ontologies collaboratives	NON	NON	OUI	OUI (en réutilisant une approche comme DILIGENT)

TABLE 4.1 – Comparaison entre les méthodologies de construction d'ontologie

des ressources textuelles et découvrir des classes, des relations de subsumption, ou des relations en se basant sur une analyse syntaxique et sur la présence de co-occurences comme Text-To-Onto [31] ou [57]. D'autres approches exploitent des informations semi-structurées, des documents en langage naturel, mais également des ontologies génériques existantes, comme [26] qui est un système d'acquisition semi-automatique d'ontologies basées sur les données semi-structurées d'un intranet d'entreprise, et des documents en langage naturel et des ontologies telles que GermaNet [10] et WordNet [11].

De plus, il existe des stratégies de modélisation [15] qui varient en fonction de l'ordre dans lequel les concepts et les relations sont considérés. Une méthodologie bottom-up démarre par les concepts les plus spécifiques et s'intéresse ensuite aux concepts de plus en plus généraux. A l'opposé, une méthodologie *top-down* démarre par les concepts les plus généraux et les spécialise au fur et à mesure. Enfin, une méthodologie *middle-out* commence cette fois par les concepts les plus importants et les généralise et/ou les spécialise.

Parmi les méthodologies de construction d'ontologie, nous trouvons METHONTOLOGY [16], On-To-Knowledge [51], DILIGENT [45] et NeOn [21] qui font partie des méthodologies les plus référencées dans ce domaine. Une comparaison entre ces méthodologies est résumée dans le tableau 4.1.

METHONTOLOGY [16] propose une méthodologie qui est composée des éléments suivants : le processus de développement de l'ontologie, un cycle de vie basé sur des prototypes évolutifs qui résultent du processus de développement, et enfin les étapes de réalisation de chaque activité, les techniques utilisées, les attendus et les méthodes d'évaluation de l'ontologie.

Le processus de développement de l'ontologie METHONTOLOGY [16] comporte les quatre phases suivantes :

— *La phase de spécification* : il s'agit d'identifier l'ensemble des termes à représenter, et les scénarios d'utilisation de l'ontologie et son expressivité.

— *La conceptualisation* : durant cette phase, il s'agit de classer les termes comme des concepts, des instances, des attributs, etc. Dans certains cas, quand il existe un nombre important de termes, [30] propose de construire une hiérarchie de concepts en utilisant des relations telles que "*sous-classe-de*" en considérant qu'une classe C est une sous-classe de la classe parente P si et seulement si chaque instance de C est

10. GermanNet : https://uni-tuebingen.de/fakultaeten/philosophische-fakultaet/fachbereiche/neuphilologie/seminar-fuer-sprachwissenschaft/arbeitsbereiche/allg-sprachwissenschaft-computerlinguistik/ressourcen/lexica/germanet-1/
11. WordNet : https://wordnet.princeton.edu/

également une instance de P. Ils définissent aussi une partition de sous-classe de C comme un ensemble de sous-classes de C qui sont mutuellement disjointes. Enfin, ils identifient également des ensembles de classes mutuellement disjointes qui couvrent complètement C (c.-à-d. chaque instance de C est une instance d'exactement une des sous-classes de la partition).

— *La phase d'intégration* : l'objectif de cette phase est d'aider l'ontologue à réutiliser les termes d'une ontologie existante en s'assurant que sa sémantique soit cohérente avec les termes identifiés dans la phase de conceptualisation.

— *La phase d'implémentation* : c'est la dernière phase de mise en œuvre de l'ontologie, où il s'agit de transformer le modèle conceptuel en un modèle implémenté. Cette transformation peut être établie dans environnement qui fournit :

- Un analyseur lexical et syntaxique pour garantir l'absence d'erreurs lexicales et syntaxiques.
- Un traducteur pour garantir la portabilité des définitions vers d'autres langues cibles.
- Un éditeur pour ajouter, supprimer ou modifier des définitions.
- Un navigateur pour inspecter la bibliothèque d'ontologies et leurs définitions.
- Un chercheur pour rechercher les définitions les plus appropriées.
- Un évaluateur pour détecter les incomplétudes, les incohérences et les connaissances redondantes.
- Un mainteneur automatique pour gérer l'inclusion, la suppression ou la modification des définitions existantes.

Bien que METHONTOLOGY permet la réutilisation des termes d'autres ontologies, cette réutilisation est limitée aux classes et ne permet pas de bénéficier des relations et des propriétés définies dans d'autres ontologies.

On-To-Knowledge [51] permet aux ontologues d'explorer et de réutiliser d'autres ontologies. On-To-Knowledge propose le processus suivant :

— *La phase d'étude de faisabilité* : le but de cette première étape est d'identifier les problématiques et les solutions potentielles, pour assurer la bonne intégration et l'exploitation de l'ontologie dans le système.

— *La phase de lancement* : Le résultat de cette étape est un document spécifiant les différentes exigences. Ce document doit décrire le but de l'ontologie, son domaine, les applications supportées par l'ontologie, les sources de connaissances, une liste des utilisateurs, les questions de compétence qui est un aperçu des requêtes possibles sur le système, et les ontologies potentiellement réutilisables durant la construction de cette nouvelle ontologie.

— *La phase de raffinement* : l'objectif de cette phase est de produire une ontologie qui satisfait les spécifications décrites dans la phase précédente. Cette phase de raffinement est composée de trois étapes :

- L'étape de la collecte des concepts pertinents donnés lors de la phase de lancement. Cette collection est appelée "taxonomie de base".
- L'étape de développement d'une "ontologie de base" qui contient les concepts pertinents, les relations

entre ces concepts et les axiomes. Ces relations sont créées via un processus d'élicitation des connaissances basé sur les données initiales de la taxonomie de base avec l'aide des experts du domaine.

- L'étape de conceptualisation et de formalisation pour exprimer l'ontologie de base avec des langages de représentation formels afin d'obtenir une "ontologie cible".

— *La phase d'évaluation* : Durant cette phase le développeur d'ontologie commence d'abord par vérifier si l'ontologie cible répond aux critères du document de spécification des exigences d'ontologie. Ensuite, l'ontologie cible est testée dans une plateforme de test par des utilisateurs bêta pour obtenir des retours d'information (feedbacks) utilisés pour améliorer l'ontologie.

— *La phase de maintenance* : Le but de cette phase est d'adapter l'ontologie aux changements qui peuvent apparaître. Différentes règles sont proposées pour contrôler le processus de mise à jour des ontologies. L'approche recommande de rassembler les modifications à apporter, de créer une nouvelle version de l'ontologie, puis de la tester avant de l'intégrer au système.

Contrairement à METHONTOLOGY et On-To-Knowledge, DILIGENT [45] est une méthodologie de construction collaborative d'ontologie. DILIGENT se concentre sur l'ingénierie des ontologies collaboratives et distribuées. Cette méthodologie permet à plusieurs développeurs d'ontologies appartenant à des organisations différentes de travailler sur la même ontologie. Cette méthodologie comporte cinq phases principales :

— *La phase de construction* : durant cette première étape, une équipe restreinte commence par produire une ontologie initiale partagée. Cette équipe est constituée des experts de domaine, des utilisateurs et des ontologues qui appartiennent à des organisations différentes.

— *La phase d'adaptation locale* : Après la production de l'ontologie initiale, les utilisateurs peuvent l'adapter localement à leurs propres besoins et sauvegarder une nouvelle version de l'ontologie -appelée locale- dans leur environnement.

— *La phase d'analyse* : L'ontologie locale est analysée par un conseil pour décider quels changements seront introduits dans la prochaine version de l'ontologie partagée.

— *La phase de révision* : Cette étape se réalise régulièrement pour réviser l'ontologie partagée, afin que les ontologies locales ne s'éloignent pas trop de l'ontologie partagée.

— *La phase de mise à jour locale* : Cette étape s'applique lorsqu'une nouvelle version de l'ontologie partagée est publiée, dans ce cas les utilisateurs peuvent mettre à jour leurs propres ontologies locales pour mieux utiliser les connaissances représentées dans la nouvelle version.

On-To-Knowledge et DILIGENT permettent la réutilisation de ressources ontologique. Néanmoins, ces méthodologies ne précisent pas les étapes nécessaires pour permettre la réutilisation et la réingénierie des ressources de connaissances existantes. NeOn [21] a été créé pour compléter les approches existantes. NeOn définit neuf scénarios qui peuvent survenir durant la construction d'une ontologie. L'utilisation de ces scénarios dépend des propriétés de l'ontologie à développer et de la méthode de développement, c.-à-d. le développement d'ontologies

collaboratives, impliquant ou non la réutilisation de ressources ou d'autres ontologies.

Les scénarios de NeOn sont les suivants :

1. De la spécification à l'implémentation : est le scénario dans lequel les développeurs construisent une ontologie à partir de zéro sans aucune réutilisation d'autres ontologies. Ils doivent dans ce cas commencer par collecter les connaissances nécessaires pour construire le document de spécification de l'ontologie. Ensuite, les développeurs doivent conceptualiser, formaliser et implémenter l'ontologie en utilisant METHONTOLOGY ou On-To-Knowledge.
2. Réutilisation et réingénierie des ressources non ontologiques : représente le scénario où les développeurs d'ontologie sélectionnent les ressources non ontologiques qui satisfont les objectifs définis dans le document de spécification. Puis, ils effectuent un processus de transformation de ces ressources pour les intégrer dans l'ontologie.
3. Réutilisation des ressources ontologiques : est le scénario qui définit trois façons de réutiliser des ontologies disponibles, qui doivent être guidées par le document de spécification :
   - Réutiliser toute l'ontologie.
   - Réutiliser seulement un module identifié représentant une sous-partie de l'ontologie.
   - Réutiliser certains triplets décrivant le niveau conceptuel de l'ontologie.
4. Réutilisation et réingénierie des ressources ontologiques : est le scénario qui concerne les ressources ontologiques. Si le besoin d'adaptation à de nouveaux besoins est exprimé, les développeurs doivent les reformaliser.
5. Réutilisation et fusion des ressources ontologiques : dans le cas où plusieurs ressources ontologiques peuvent être réutilisées, les développeurs d'ontologie peuvent les fusionner pour créer une nouvelle ontologie, ou établir des alignements entre les ontologies afin d'en créer une nouvelle.
6. Réutilisation, fusion et réingénierie des ressources ontologiques : Ce scénario a la même séquence d'activités que le scénario précédent, cependant ici, les développeurs d'ontologie peuvent décider de ne pas utiliser l'ensemble des ressources ontologiques fusionnées tel quel, mais de le remanier.
7. Réutilisation de modèles de conception d'ontologies : représente le scénario où les développeurs d'ontologies peuvent accéder aux référentiels de modèles de conception d'ontologie [12] pour les réutiliser afin de faciliter et accélérer la modélisation. Les modèles de conception d'ontologie peuvent être génériques ou dédiés à un domaine particulier comme l'agriculture, la biologie, la pêche, etc.
8. Restructuration des ressources ontologiques : ce scénario décrit la restructuration des ressources ontologiques avant de les intégrer dans l'ontologie en cours de construction. Il peut s'agir de [21] :

12. Les modèles de conception d'ontologie : `http://ontologydesignpatterns.org/wiki/Main_Page`

- Modulariser l'ontologie en créant différentes sous-parties d'ontologie.
- Élaguer les branches de la taxonomie qui ne sont pas nécessaires.
- Inclure de nouveaux concepts et relations à l'ontologie.
- Spécialiser certaines branches en incluant des concepts et des relations de domaine plus précis.

9. Localiser les ressources ontologiques : les développeurs d'ontologies adaptent dans ce scénario une ontologie existante en d'autres langues et en d'autres cultures en traduisant les labels de l'ontologie dans une ou plusieurs langues afin d'obtenir une ontologie multilingue.

Chacun de ces neuf scénarios peut être combiné avec un autre selon les exigences de la situation, et chacune de ces combinaisons doit inclure les activités de base du premier scénario qui doit être effectué dans tout le processus de développement d'ontologie.

4.6 Ontologie ASBESTOS

4.6.1 Méthodologie de construction utilisée

Nous nous sommes basés sur un ensemble de scénarios décrits dans la méthodologie NeOn pour construire l'ontologie Asbestos permettant de décrire les bâtiments, leurs propriétés et les diagnostics associés. Il s'agit principalement des scénarios permettant la réutilisation de ressources ontologiques et non ontologiques existantes. Plus précisément, nous avons utilisé les scénarios suivants :

— Le scénario 1 de la spécification à l'implémentation : pour créer les parties (modules) de l'ontologie qui n'existent pas dans d'autres ressources.

— Le scénario 2 de la réutilisation et réingénierie des ressources non ontologiques : afin d'utiliser les documents du CSTB ou des ressources documentaires externes qui contiennent des données tabulaires.

— Le scénario 4 de la réutilisation et réingénierie des ressources ontologiques. En effet, nous avons adapté et simplifié la représentation des intervalles de temps et des instants décrits dans la Time Ontology (OWL-Time) présentée dans la Section 4.4.2.

Les étapes principales pour construire notre base de connaissance sont :

1. Construire une ontologie Amiante qui permet de modéliser les connaissances sur les bâtiments et les diagnostics réalisés sur les bâtiments quand ils existent. Dans cette étape, nous construisons la partie haute de l'ontologie qui contient le modèle standard suivi par tous les bâtiments et défini dans la norme NF X46-020. Ce modèle contient les différents composants d'un bâtiment et les relations entre eux.

2. Enrichir l'ontologie par de nouveaux concepts et sous-concepts issus d'un processus d'extraction semi-automatique des informations décrites dans les ressources documentaires du CSTB. c.-à-d. si un concept

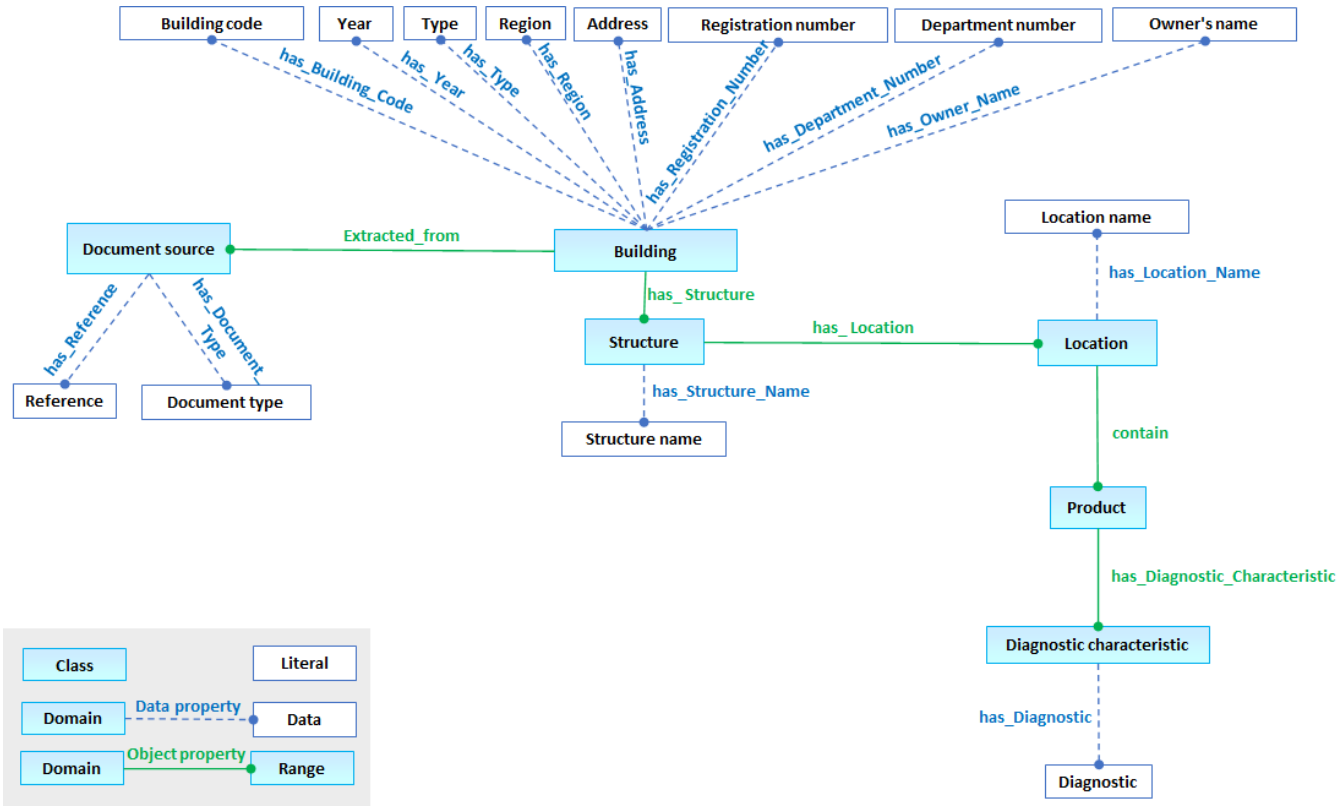


FIGURE 4.5 – Concepts principaux de l'ontologie Amiante

n'était pas motionné dans la version actuelle de l'ontologie, le processus de l'enrichissement va mettre à jour cette version et va intégrer le nouveau concept dans la structure de l'ontologie.

3. Peupler l'ontologie avec les instances de concepts et de propriétés issues d'un processus d'extraction automatique des informations décrites dans les ressources documentaires du CSTB.

4.6.2 Partie haute de l'ontologie ASBESTOS

Dans cette section, nous présentons la partie haute de l'ontologie ASBESTOS (c.f. Figure 4.5) qui a été construite en exploitant les documents du CSTB, les connaissances des experts, et les besoins de prédiction dans le projet ORIGAMI ([32]).

Les principaux concepts de cette ontologie sont les suivants :

— Building : construction caractérisée par un code CSTB qui correspond à un type de bâtiment donné, le type de bâtiment (p. ex. école, maison, etc.), l'année de construction et l'adresse du bâtiment.

— Structure : espace faisant partie du bâtiment (p. ex. balcon, toit, escalier, etc.).

— Location : localisation composant une structure (p. ex. porte, fenêtre, mur, etc.).

— Product : produit utilisé dans une localisation (p. ex. colle, enduit, etc.).

— Document Source : Un document source est décrit par son type (projet type homologué ou diagnostic

amiante) et son url vers le fichier source.

— Diagnostic characteristic : qui est composé des informations extraites à partir des diagnostics. Il contient le résultat de l'existence d'amiante dans le produit (si l'information de diagnostic existe). Ce concept est décrit par le résultat de l'existence d'amiante propriété "has_Diagnostic" qui prend la valeur 0 ou 1.

L'ontologie ASBESTOS a été enrichie par 8 sous-classes de structures, 19 sous-classes de localisation et 38 sous-classes de produit. La partie haute de l'ontologie est construite en anglais, cependant nous avons gardé les mêmes labels des individus dans les documents de CSTB qui sont en français.

4.7 Enrichissement et peuplement de l'ontologie

Les deux procédures qui vont être représentées dans cette section sont l'enrichissement et le peuplement de l'ontologie. Le peuplement est l'ajout des individus avec ses propriétés (Data Property) et les relations qui les lient aux autres individus (Object Property). Ces individus sont extraits à partir des documents CSV. Nous détaillons les procédures de l'extraction des données dans la section 4.8.

Chaque individu est une instance d'un concept précis dans l'ontologie, si nous trouvons un individu qui appartient à un concept qui n'existe pas dans l'ontologie, nous enrichissons d'abord l'ontologie avec ce concept puis nous ajoutons cet individu à l'ontologie. L'enrichissement est la phase d'ajout des nouveaux concepts (ou des sous-concepts) à l'ontologie de base, par exemple lorsque nous trouvons un individu appartenant à un nouveau concept "enduit" qui n'existe pas dans l'ontologie de base, sachant que "enduit" est un sous-concept de "produit", nous l'ajoutons à l'ontologie comme un sous-concept de "produit".

A la fin de la procédure de l'enrichissement avec les différentes données, l'ontologie ASBESTOS contient 139 concepts, dont 8 sont des sous-concepts de "Structure", 20 sont des sous-concepts de "Localisation" et 102 sont des sous-concepts de "Produit". Parmi les sous-concepts de "Produit" 59 sont des sous-concepts directs. La Figure 4.6 montre un extrait de l'ensemble de concepts de l'ontologie ASBESTOS après l'enrichissement.

Après le peuplement l'ontologie contient 51970 triplets qui décrivent 2998 instances de produit, 341 localisations, 214 structures et 94 bâtiments.

4.8 Extraction automatique des données

Durant l'enrichissement et le peuplement de l'ontologie, nous traitons des données tabulaires et textuelles dans des documents différents (diagnostics et projets homologués). La représentation de ces données se diffère d'un document à l'autre d'où la nécessité d'une procédure d'extraction qui emploie différentes techniques pour reconnaître et classer les données (produit, localisation, structure, etc.) et détecter les relations les liants.

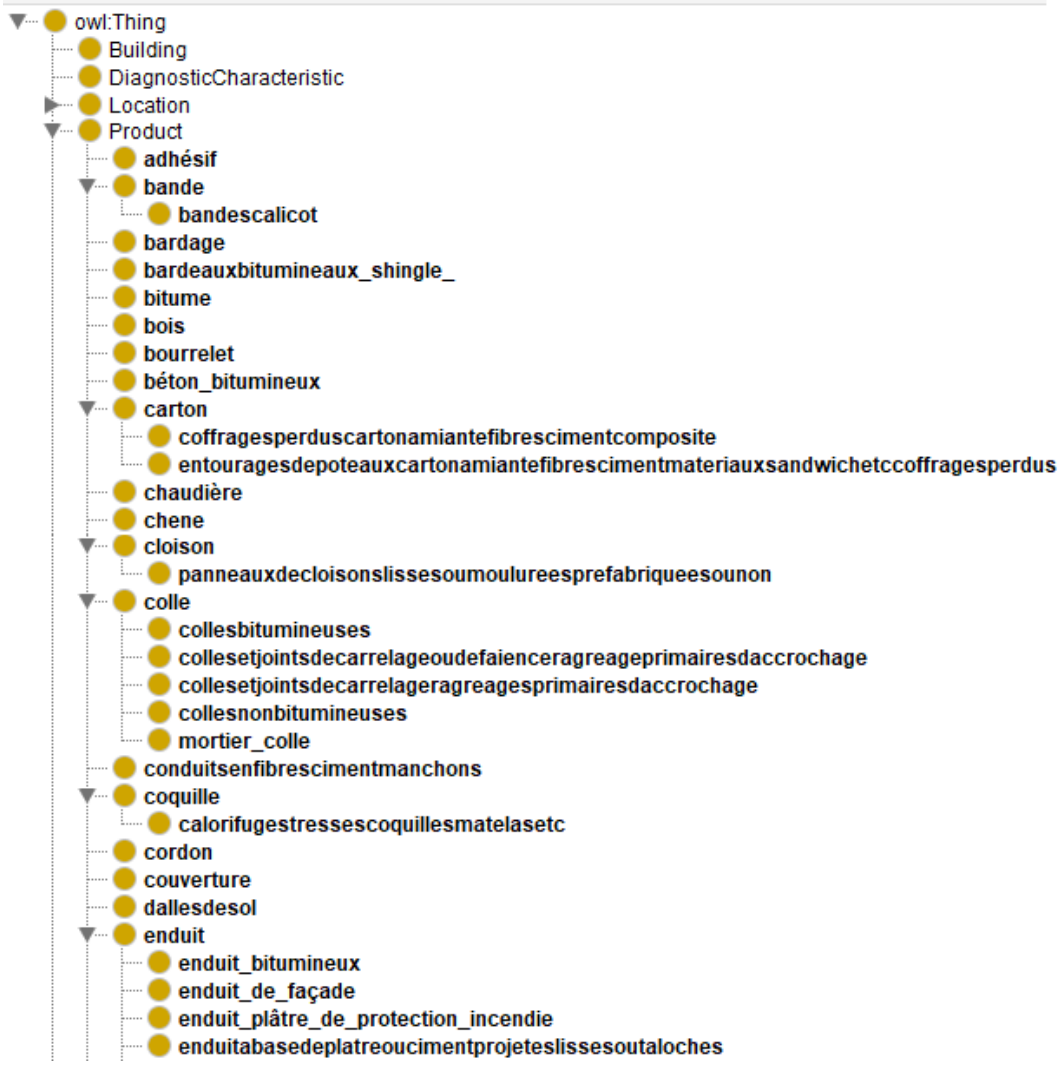


FIGURE 4.6 – Extrait de l'ensemble de concepts de l'ontologie ASBESTOS visualisé avec Protégé après l'enrichissement

4.8.1 Extraction des données depuis les diagnostics

Les diagnostics sont construits sous la forme des tableaux dont chaque colonne contient un composant précis (voir la Figure 4.1), par exemple la première colonne contient les identifiants des bâtiments. Grâce à cette représentation nous avons le type de chaque individu c.-à-d. les individus de la première colonne sont des bâtiments, de la deuxième sont des produits, etc. Les individus dans la même ligne possèdent des relations existentielles entre eux, c.-à-d. le produit de la première ligne existe dans la localisation de la première ligne qui existe dans la structure de la première ligne qui elle aussi existe dans le bâtiment de la même ligne. Le nom d'un individu est composé du nom du type de l'individu concaténé avec un identifiant incrémental. Alors l'extraction depuis la première ligne, par exemple, résulte de la création d'un individu "bandecalicot_1" de type "bande calicot" qui est un sous-concept de "produit", un individu "cloisonseche_1" de type "cloison sèche" qui est un sous-concept de "location", un individu "paroiverticalintérieur_1" de type "paroi vertical intérieur" qui est un sous-concept de "structure" et un individu "bat_111002"

de type “building”. Ensuite “bandecalicot_1” sera relié avec “cloisonseche_1” via la propriété d’objet “contain”, “cloisonseche_1” avec “paroiverticalintérieur_1” via “has_Location” et “paroiverticalintérieur_1” avec “bat_111002” via “has_Structure”. Nous ajoutons ensuite les propriétés de données pour le bâtiment : l’année 1986 avec la propriété “has_Year” et la région “Palaiseau” avec la propriété “has_Region”. Chaque individu de produit est associé à une instance de “Diagnostic characteristic” où nous l’associons avec son diagnostic (1 dans le cas de la première ligne) via la propriété “has_Diagnostic”.

4.8.2 Extraction des données depuis les projets types homologués

Les projets homologués sont représentés de manière différente que les diagnostics. Comme la Figure 4.3 montre, la première colonne contient le nom de la structure, mais la deuxième contient une description textuelle de la composition de cette structure (c.-à-d. l’ensemble de localisations et de produits qui composent la structure), alors nous avons besoin d’extrait ces composants pour les ajouter à l’ontologie. Ils existent plusieurs méthodes et approches qui s’intéressent à l’extraction de ces entités nommées depuis les textes. Ces approches sont utilisées dans plusieurs domaines tels que la génération d’ontologies. Cependant, les descriptions textuelles dans notre cas sont toutes sous la forme des passages déclaratifs qui suivent un style descriptif pour décrire la composition d’une structure. Les compositions des phrases utilisées suivent le registre de langue courant [13]. Parmi les caractéristiques des descriptifs, les entités sont toujours décrits par le terme défini par la norme NF X46-020 et n’utilise pas d’autres synonymes. Par exemple pour les revêtements, nous trouvons qu’ils sont toujours décrits par le terme “revêtement” et pas par “couverture” ou d’autres termes. Grâce à la simplicité des passages textuels, nous n’avions pas besoin d’une approche complexe ou très avancée pour faire l’extraction des termes, alors nous avons défini des patterns pour détecter les termes dans le texte. Afin de définir les patterns, nous avons utilisé les expressions régulières pour définir la forme générale et la composition d’un terme. En analysant les différents noms de concepts (produits et localisations), nous avons identifié l’expression régulière suivante qui couvre tous ces noms :

$$(ADJ^* \; NOM \; PRP? \; ADJ^*)^+$$

tel que :

ADJ représente un adjectif comme : “métallique”.

NOM représente un nom comme : “béton”.

PRP représente une préposition comme : “à”, “de” .etc. Mais nous excluons “sans”, “avec” et “en” qui sont considérés comme des séparateurs entre les composants, c.-à-d. si nous trouvons par exemple “carreaux de plâtre avec peinture” nous extrairons deux composants : “carreaux de plâtre” et “peinture”. Ensuite, nous ajouterons les deux composants à l’ontologie. Cependant, dans le cas de “sans” le deuxième composant ne sera pas ajouté à l’ontologie. Dans le cas de “en”, ce dernier lie une localisation avec ses produits.

13. Le registre de langue courant : `https://fr.wikipedia.org/wiki/Registres_de_langue_en_français`

Dans ces textes, il y a aussi différents séparateurs qui sont :

"-" et ";" : sont les séparateurs utilisés entre les descriptions de différentes localisations, par exemple : "Localisation1 contient produit1 - localisation2 contient produit2".

" :" : sépare entre les noms de localisations qui procèdent la même composition et ses composants, par exemple : "Localisation1 et Localisation2 : produit1 avec produit2".

La conjonction de coordination "et" : ce type de séparateurs lie les composants qui partagent la même propriété de co-existence (par exemple "contient produit1 et produit2") ou d'absence (par exemple "sans produit1 et produit2").

Afin d'identifier la nature de chaque mot dans les discours textuels, nous avons utilisé un tagueur de texte *treetaggerwrapper*[14] qui prend en charge la langue française et qui associe à chaque mot dans un texte un label comme NOM, ADJ, PRP, VER, .etc.

Prenant l'exemple du projet homologué montré dans la Figure 4.2 pour appliquer notre méthode d'identification de termes pour quelques lignes.

- Ligne A : Dans cette ligne nous avons le séparateur "en" entre la localisation et le produit (respectivement "Fouilles" et "rigoles") qui suivent la forme de NOM.
- Ligne B : après l'application de l'expression régulière sur cette ligne nous obtenons : "Murettes béton de 0,50*0,30." $\rightarrow$ nous récupérons le terme "Murettes béton" qui suive la forme ($NOM\ NOM$). Comme il y a un seul terme dans cette ligne alors il sera classé comme un produit puisque la localisation dans les projets homologués peut ne pas être mentionnée et dans ce cas elle sera représentée dans l'ontologie par un nœud blanc.
- Ligne C : "Ossature métallique et murs à double paroi : dalles ciment de 1,00*,040*0,04 et carreaux de plâtre creux lissés 2 faces, de 0,60*0,40*0,07." $\rightarrow$ il y a des " :" ce qui signifie qu'ils existent plusieurs localisations avant les " :" qui partagent la même composition qui vient après, avant les " :" il y a aussi un séparateur ("et"), après l'application de l'expression régulière nous obtenons "Ossature métallique" ($NOM\ ADJ$) et "murs à double paroi" ($NOM\ PRP\ ADJ\ NOM$). Les deux localisations vont contenir les mêmes produits suivants : "dalles ciment" ($NOM\ NOM$) et "carreaux de plâtre creux lissés" ($NOM\ PRP\ NOM\ ADJ\ ADJ$).

Il faut noter que la procédure d'extraction et de classification de termes est semi-automatique, l'expert vérifie à la fin si tous les termes sont bien classés.

4.9 Conclusion

Nous avons présenté dans ce chapitre l'ontologie ASBESTOS qui peut être considérée comme la première ontologie permettant de représenter les caractéristiques des bâtiments et les diagnostics amiante qui ont été réalisés sur des produits les constituant. L'ontologie permet également de stocker l'origine de ces informations en conservant

14. https://pypi.org/project/treetaggerwrapper/

les références des documents sources.

Le haut de l'ontologie ASBESTOS a été créé manuellement et validé par l'expert du CSTB. Cette ontologie a ensuite été enrichie semi-automatiquement en exploitant les ressources du CSTB qui utilisent un vocabulaire conforme à la norme NF X46-020. La méthodologie suivie se base principalement sur la méthodologie NeOn qui décrit des scénarios impliquant la réutilisation de sources ontologiques ou non ontologiques.

L'ontologie ASBESTOS nous a permis de représenter 2998 diagnostics ainsi que quelques projets homologués du CSTB. Elle permet ainsi aux utilisateurs du CSTB de requêter les diagnostics et d'accéder à la source documentaire de chaque donnée sauvegardée. Cette ontologie a été mise à disposition du CSTB pour l'utiliser pour les données concernant l'amiante, mais aussi dans d'autres projets qui impliquent des descriptions de bâtiments et qui concernent d'autres domaines comme la présence de plomb ou le ré-emploi de produits et matériaux dans le cadre de l'économie circulaire.

Chapitre 5

Calcul de probabilité de présence d'amiante dans les parties de bâtiments basé sur des ressources externes

5.1 Introduction

Dans ce chapitre, nous présentons la première approche semi-supervisée pour la prédiction de l'amiante basée sur notre ontologie ASBESTOS.

Dans les documents du CSTB, on ne connaît que la classe de produit utilisée pour une localisation, mais pas la référence du produit utilisé. Par exemple, on sait que le produit utilisé est un *Enduit,* mais on ne sait pas quel enduit (c.-à-d. quel produit commercialisé) est exactement utilisé. Il existe des ressources externes décrivant des listes de produits commercialisés qui ont été amiantés à certaines périodes, mais il n'est pas possible de lier directement cette connaissance aux descriptions des produits utilisés dans les bâtiments. Dans ce qui suit, nous proposons d'utiliser une approche qualifiée d'hybride [33, 32] basée sur la présence d'amiante dans les produits commercialisés pour prédire la probabilité d'existence d'amiante pour les produits d'un bâtiment construit à une date donnée, puis plus généralement pour ses localisations, ses structures, et le bâtiment lui-même.

Cette approche calcule d'abord une probabilité de présence d'amiante pour un sous-ensemble de produits puis elle peut propager la probabilité obtenue via des règles exprimées en SWRL qui utilisent la classe de produit, l'année de construction du bâtiment.

Nous décrivons dans la première section les ressources externes utilisées par l'approche hybride (Section 5.2). Nous présentons ensuite dans la Section 5.3 le module d'enrichissement et de peuplement de l'ontologie ASBESTOS avec les données de ressources externes, où nous montrons la procédure d'extraction de données et la

résolution des contradictions lorsque nous devons faire face à des données conflictuelles. Dans la Section 5.4, nous présentons comment la probabilité de présence d'amiante dans les produits est calculée en utilisant les ressources externes, et comment elle sera ajustée avec un sous-ensemble de diagnostics amiante. Les règles et l'approche de propagation sont présentées dans les deux sections suivantes (Sections 5.6 et 5.5). Nous clôturons ce chapitre en présentant les expérimentations (Sections 5.7) réalisées sur un sous-ensemble de bâtiments du CSTB. Plus précisément, nous détaillerons une analyse quantitative et qualitative des résultats et un ensemble de comparaison avec d'autres approches sur le même échantillon de données. Nous montrerons également que cette approche basée sur des ressources externes peut obtenir de bons résultats même si l'échantillon de donnée est de petite taille.

5.2 Description des ressources externes

Nous utilisons deux ressources externes fournies par l'ANDEVA [1] et INRS [2].

— L'ANDEVA publie sur son site Internet une liste de 650 produits commercialisés contenant de l'amiante. Cette liste contient le nom du produit et l'année à partir de laquelle il ne contient plus d'amiante. Le tableau 5.2 montre l'exemple de trois produits commercialisés décrits dans cette ressource. La colonne "Fournisseur et produit" contient le nom du fournisseur s'il existe suivie par le nom du produit commercialisé séparés par une "," (e.g. TREMCO, MONO) tandis que la colonne "Nom de famille de produits" contient le nom de la classe de ce produit (e.g. Mastic) en utilisant le vocabulaire utilisé dans la norme NF X46-020. La colonne "Amianté jusqu'en" contient la dernière année où ce produit a été amianté, à partir de cette année, l'ANDEVA suppose que le produit devient non-amianté.

— L'INRS publie également une liste de 300 produits qui décrit pour un ensemble de noms de produits, les intervalles de temps pendant lesquels ces produits ont contenu de l'amiante avec une éventuelle incertitude sur certains intervalles de temps. Le tableau 5.1 montre l'exemple de trois produits commercialisés décrits par l'INRS où nous trouvons le nom commercial du produit dans la colonne "Produit" (p. ex., ARMAZOL), la classe de chaque produit est mentionnée dans la colonne "Type d'utilisation" conformément à la norme NF X46-020 (p. ex., Revêtements de sols en dalles ou en rouleaux), l'information de présence d'amiante est située dans la colonne "Renseignements divers" (p. ex., l'ARMAZOL a été amianté jusqu'en 1982, mais non renseignée après). L'INRS mentionne aussi parfois d'autres informations comme le nom de fournisseur et le type d'amiante (i.e. Amiante, Chrysotile).

Les deux listes de l'ANDEVA et l'INRS sont représentées sous forme des tableaux dans des fichiers .csv où les colonnes sont séparées par des ";".

1. Association nationale de défense des victimes de l'amiante http://andeva.free.fr/expositions/gt_expos_produits.htm
2. Institut National de la Recherche et de la Sécurité http://www.inrs.fr/media.html?refINRS=ED%201475

Produit	Fournisseur	Renseignements divers	Type d'amiante	Type d'utilisation
ARMAZOL		Amianté jusqu'en 1982, non renseigné après	Amiante	Revêtement de sols en dalles ou en rouleaux
CALITHAN	EMFI	Amianté jusqu'en 1994, sans amiante après	Chrysotile	Colle
MONO	TREMCO	Amianté jusqu'en 1989, sans amiante après	Chrysotile	Mastic

TABLE 5.1 – Extrait de l'INRS : "ARMAZOL"

Nom de famille de produits	Fournisseur et produit	Amianté jusqu'en
Revêtements de sols en dalles ou en rouleaux	ARMAZOL	1990
Colle	EMFI, CALITHAN	1994
Mastic	TREMCO, MONO	1995

TABLE 5.2 – Extrait de l'ANDEVA : "ARMAZOL"

5.3 Enrichissement et peuplement d'ontologie avec des produits commercialisés

Dans cette section, nous présentons comment les deux ressources externes fournies par l'ANDEVA et l'INRS ont été utilisées pour enrichir et peupler l'ontologie ASBESTOS avec des produits commercialisés.

5.3.1 Extraction automatique des classes de produits et des descriptions de produits dans les données tabulaires

Nous avons enrichi l'ontologie Asbestos avec les classes de produits mentionnées dans les ressources ANDEVA et INRS. Le vocabulaire est uniforme car basé sur la norme NF X46-020, mais la norme n'indique pas les relations hiérarchiques entre classes. Toutes les classes extraites sont des sous-classes de la classe $Product$ et les relations de subsomption sont créées en se basant simplement sur l'inclusion de chaînes de caractères. Par exemple, la classe "Enduit de façade" et une sous-classe de "Enduit".

L'objectif est de représenter les caractéristiques concernant l'amiante d'un produit tel qu'il est décrit dans chaque source, ainsi que les caractéristiques résultant de la fusion de ces informations afin de les utiliser dans l'approche hybride tout en conservant leur provenance. Pour représenter ces informations, nous avons ajouté à la partie haute de l'ontologie un nouveau module dont les principaux concepts sont $Extracted\ characteristic$ et $Calculated\ characteristic$. $Extracted\ characteristic$ contient les données extraites et le résultat de la fusion de ces données. $Calculated\ characteristic$ permet la représentation des résultats obtenus par l'approche hybride pour un produit présent dans un bâtiment donné.

Plus précisément, $Extracted\ characteristic$ nous permet de représenter l'intervalle de temps, la probabilité amiante, et la ressource dont est issue cette caractéristique (c.-à-d. data propriété has_Source ayant pour valeur

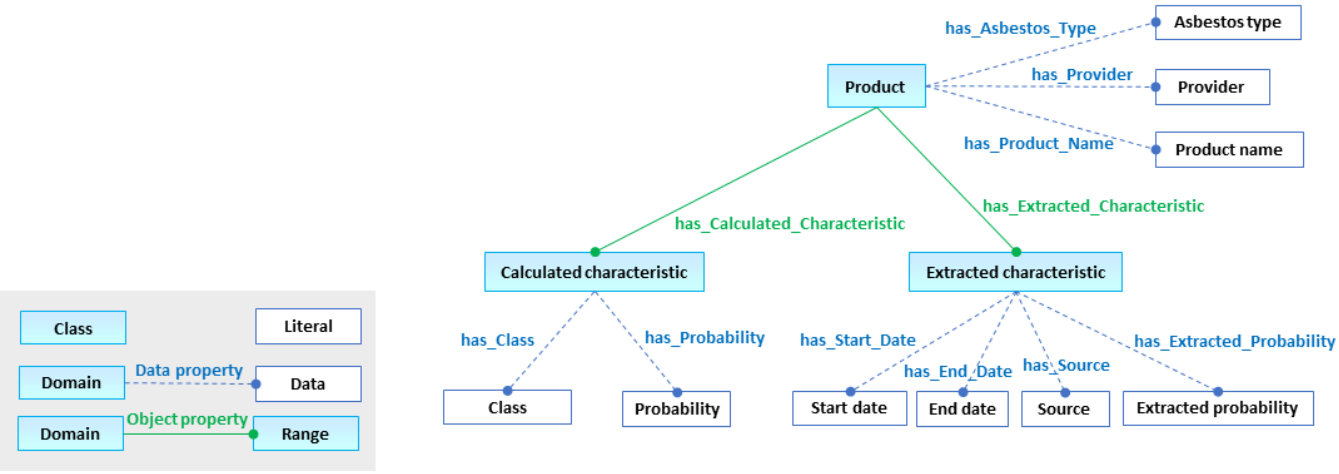


FIGURE 5.1 – Les concepts du module de l'approche hybride

INRS, ANDEVA ou FUSION). L'approche utilise une probabilité de 1 pour les produits contenant de l'amiante sur la période, 0,5 lorsque la présence d'amiante est inconnue et 0 sinon (c.-à-d. sans amiante). Ces probabilités sont stockées dans la propriété $has\_Extracted\_Probability$. Étant donné que les deux ressources externes sont également fiables, nous avons appliqué une approche pessimiste qui considère pour chaque intervalle de temps la probabilité la plus élevée d'amiante dans le processus de fusion.

La figure 5.1 montre les concepts et les propriétés ajoutées à l'ontologie ASBESTOS pour manipuler l'approche hybride :

— Product name : pour mettre le nom du produit commercialisé.

— Provider : représente le nom du fournisseur.

— Asbestos type : pour mettre le type d'amiante utilisé dans le produit commercialisé (Chrysotile ou Amiante).

Ainsi, nous avons ajouté des propriétés au nouveau concept $Extracted\ characteristic$:

— has_Source : pour mettre la source de l'information, si elle est extraite à partir de l'ANDEVA, de l'INRS ou de la procédure de la fusion "FUSION".

— $has\_Extracted\_Probability$: contient une des trois valeurs possibles de probabilité de présence d'amiante selon la source (0, 1 ou 0,5).

— Pour représenter l'intervalle de temps, nous nous sommes inspirés de l'ontologie du temps (OWL-Time). Nous avons identifié un intervalle de temps par deux instants temporels ; un instant de début has_Start_Date et un instant de fin has_End_Date. Lorsqu'un produit commercialisé est amianté selon l'INRS par exemple depuis 1960 jusqu'à 1970, nous mettons dans $Source$ INRS, dans $Extracted\ probability$ la valeur 1, dans has_Start_Date la date de début de l'intervalle 1960 et dans has_End_Date la date de fin 1970.

Prenons l'exemple de l'"ARMAZOL" qui est décrit dans l'INRS et l'ANDEVA dans les tableaux 5.1 et 5.2 respectivement, et qui ne comporte pas de mention de fournisseur.

A l'issue de cette étape, la classe de produit "Revêtements de sols en dalles ou en rouleaux" est créée si elle

n'existe pas déjà et une instance de cette classe dont le nom est "ARMAZOL" en lien avec 6 instances de la classe "Extracted characteristic" sont créées. Ils ont les caractéristiques suivantes :

1. La première propriété extraite est présentée comme provenant de la source INRS, et comme étant amianté (probabilité = 1) de 1946 à 1982,
2. Puis comme potentiellement amianté de 1983 jusqu'en 1997 (probabilité = 0,5),
3. Les deux propriétés suivantes sont extraites est présentée comme provenant de la source ANDEVA, la première est amiantée de 1946 à 1990 (probabilité = 1),
4. Puis non amianté (probabilité = 0) de 1991 jusqu'en 1997,
5. Les deux dernières résultent de la fusion, la première le considère comme amianté de 1946 jusqu'en 1990 (probabilité = 1),
6. Et l'autre comme potentiellement amianté de 1991 jusqu'en 1997 (probabilité = 0,5).

5.3.2 Fusion des descriptions de produits identiques

Dans un deuxième temps, nous fusionnons les descriptions de produits identiques. Pour cela, nous devons décider que ces descriptions réfèrent bien au même produit (c.-à-d. liage de données) puis résoudre les éventuels conflits quand les sources de données ne s'accordent pas sur certaines valeurs de propriété.

L'étape de liage de données s'effectue simplement en nous basant sur le nom commercial du produit qui ne diffère pas selon les deux sources quand il s'agit du même produit (c.-à-d. sur l'égalité de chaînes de caractères). Un produit commercialisé n'est associé qu'à un fournisseur. Comme les deux ressources respectent la norme NF, les labels des classes de produits mentionnées sont toujours identiques pour le même produit commercialisé à quelques exceptions près. Ainsi, l'INRS et l'ANDEVA associent parfois le même produit à deux classes différentes, mais sémantiquement proches. Par exemple le produit commercialisé "FILGUM" est considéré comme un "Enduit" par l'ANDEVA et comme un "Mastic" par l'INRS. Dans ce cas, nous associons l'instance de "FILGUM" au deux classes de produit.

Dans le cas des intervalles correspondant aux caractéristiques amiante, les intervalles et les degrés de présence d'amiante peuvent être différents dans l'INRS et l'ANDEVA, comme c'est le cas dans l'exemple de l'ARMAZOL décrit précédemment. Aussi, nous effectuons la fusion de la manière suivante :

Nous considérons l'union ordonnée des bornes des intervalles de temps successifs de l'ANDEVA et l'INRS dans lesquelles le produit est commercialisé. Pour chaque paire successive de bornes, nous créons un intervalle de temps et associons à cet intervalle une probabilité de présence d'amiante qui correspond au degré de présence d'amiante le plus élevé des deux ressources. Les deux ressources étant de même fiabilité, nous appliquons ici une approche pessimiste qui considère le plus haut degré de présence d'amiante et c'est ce résultat qui sera utilisé dans la prédiction de présence d'amiante dans un bâtiment.

Ainsi, la fusion des caractéristiques amiante du produit ARMAZOL (schématisée en figure 5.2) conduit aux étapes suivantes :

1. Après avoir ordonné les bornes des intervalles d'entrée des deux sources, on obtient la liste de bornes ordonnées suivante : {1946, 1982, 1990, 1997}.
2. A partir de ces bornes, nous construisons les intervalles qui seront associés à une probabilité dans le résultat : {[1946, 1982[, [1982, 1990[, [1990, 1997[}.
3. Les intervalles [1982, 1990[et [1990, 1997[contiennent des informations contradictoires : probabilité = 0.5 pour l'INRS et probabilité = 1 pour l'ANDEVA pour le premier intervalle et probabilité = 0,5 et probabilité = 0 pour le deuxième intervalle. Pour résoudre ces conflits, nous prenons le maximum des deux valeurs (probabilité = 1 pour le premier intervalle et 0,5 pour le deuxième).

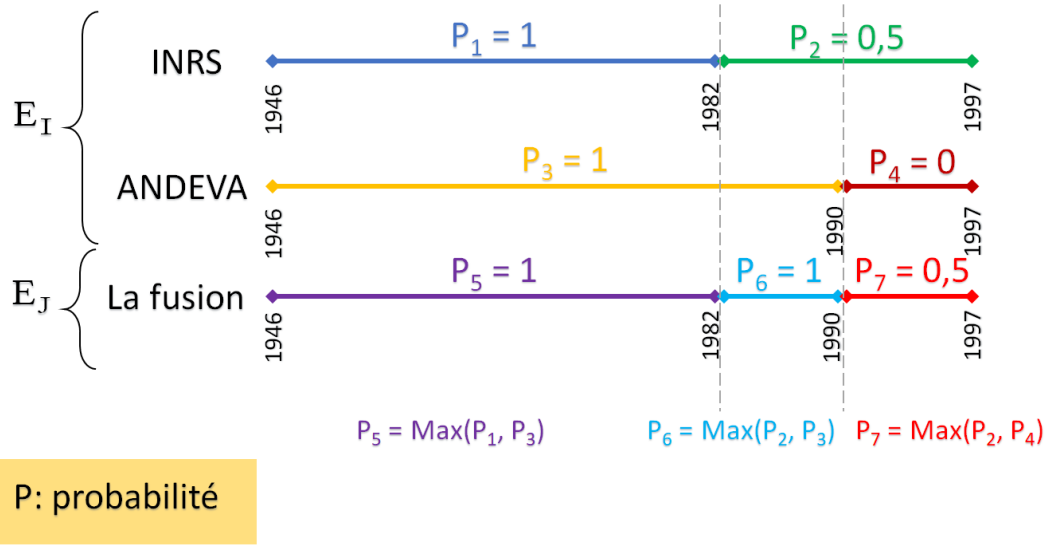


FIGURE 5.2 – Exemple de fusion des intervalles de temps et des probabilités de présence d'amiante

5.4 Calcul de la probabilité de présence d'amiante pour un produit utilisé dans un bâtiment

Comme les documents ne mentionnent pas les produits réellement utilisés lors de la construction d'un bâtiment, nous faisons l'hypothèse que nous pouvons calculer la probabilité d'existence d'amiante pour un produit utilisé à partir des produits de la même classe commercialisés au moment de la construction de ce bâtiment, en considérant qu'ils sont équiprobables. En effet, nous ne disposons pas d'information sur les parts de marché des différents produits commercialisés.

Ainsi, pour chaque produit p_k appartenant à la classe $F(p_k)$ qui est utilisé dans un bâtiment construit à une date d, la probabilité de présence d'amiante $p_a(p_k, d)$ est calculée en sommant les probabilités fusionnées p_{ext} des produits commercialisés p_j de la même classe $F(p_k)$ qui sont utilisés à cette date d et cette somme est divisée par le nombre total de produits de cette classe qui étaient en cours d'utilisation à cette date :

$$p_a(p_k, d) = \frac{\sum_{p_j \in F(p_k)} p_{ext}(p_j, d)}{\mid p_j \mid} \tag{5.1}$$

où :

$p_a(p_k, d)$ est la probabilité amiante calculée pour le produit inconnu p_k à la date d ,

$p_{ext}(p_j, d)$ est la probabilité amiante après le processus de fusion du produit commercialisé p_j et qui est de même classe que p_k ($F(p_k)$) .

Par exemple, pour calculer la probabilité de l'existence d'amiante dans un produit de la classe de produits "Peinture" en 1994, nous allons utiliser les informations fusionnées des deux ressources externes pour appliquer notre formule 5.1. La classe de produits "Peinture" contient les quatre produits décrits dans le tableau 5.3 qui reporte uniquement les intervalles de temps contenant l'année 1994 avec la probabilité de présence d'amiante associée.

Le tableau 5.3 montre un échantillon de 4 produits commercialisés de cette classe, pour lesquels les probabilités suivantes ont été générées lors de la fusion :

— "PEINTURES 54 S" n'est plus amianté à partir de 1992, donc sa probabilité en 1994 est 0.

— Dans le cas de "FILLER COAT BINDER 900-15" et "FLEXIROC HURON I RH 168" où l'INRS et l'ANDEVA ne disposaient pas d'informations sur la présence d'amiante, la probabilité à 0,5 a été générée.

— "BITUSEALAC EG" n'est plus commercialisé à partir de 1984 alors il ne sera pas pris en compte dans le calcule de la probabilité en 1994 et donc il y a que 3 produits commercialisés en 1994 de la classe "Peinture".

La probabilité de présence d'amiante pour "Peinture" est alors calculée pour $d = 1994$, $p_k = Peinture$ et un nombre de produits $\mid p_j \mid = 3$:

$$p_a(Peinture, 1994) = \frac{\sum_{p_j \in F(Peinture)} p_{ext}(p_j, 1994)}{3} = \frac{0 + 0,5 + 0,5}{3} = 0,33$$

Cependant, l'une des difficultés est que l'INRS et l'ANDEVA se focalisent uniquement sur les produits ayant été amiantés ou pour lesquels il existe une suspicion de présence d'amiante pendant au moins une période durant leur commercialisation. Ne disposant pas du nombre réel de produits commercialisés à une période donnée, nous estimons que le nombre de produits commercialisés total est largement sous-estimé et ce nombre peut varier en fonction des années et de la classe de produit. Nous proposons de le réajuster en se basant sur l'ensemble des

Produit	Intervalle fusionné	Probabilité fusionnée
PEINTURES 54 S	[1992, 1997[	0
FILLER COAT BINDER 900-15	[1985, 1997[	0,5
FLEXIROC HURON I RH 168	[1986, 1997[	0,5
BITUSEALAC EG	[1984, 1997[	Commercialisation abandonnée

TABLE 5.3 – Probabilités fusionnées de présence d'amiante des produits de la classe "Peinture" pour les intervalles de temps contenant l'année 1994

diagnostics de prélèvement disponibles. Nous comparons pour cela, pour une année donnée d et une classe de produit $F(p_k)$, la proportion de produits amiantés dans les ressources externes par rapport à l'ensemble des produits commercialisés issus des ressources externes, avec la proportion de produits amiantés dans les diagnostics réalisés pour cette même année d et même classe de produit $F(p_k)$ par rapport à l'ensemble des diagnostics posés pour l'année d et la classe $F(p_k)$. Cela nous permet de déterminer quelle est la proportion α de produits commercialisés manquants que nous considérons comme étant non amiantés.

$$p'_a(p_k, d) = \frac{\sum_{p_j \in F(p_k)} p_{ext}(p_j, d)}{\mid p_j \mid + (\alpha \times \mid p_j \mid)} \tag{5.2}$$

Nous notons par p'_a la probabilité réajustée en utilisant un ensemble de diagnostics existants :

$$p'_a(p_k, d) = p_a(p_k, d) \times (1 + \omega) \tag{5.3}$$

Dans l'équation 5.3, ω représente la différence entre la probabilité calculée et la probabilité réelle. Il résulte de la comparaison du ratio de produits amiantés dans les diagnostics avec le ratio de produits amiantés dans les ressources externes (INRS et ANDEVA) de la même année. Il est calculé en fonction de la classe de produit et l'année de construction du bâtiment comme suit :

$$\omega = \frac{\mid p_{diag,a} \mid}{\mid p_{diag} \mid} - \frac{\sum_{p_j \in F(p_k)} p_a(p_j, d)}{\mid p_j \mid} \tag{5.4}$$

où :

$\mid p_{diag,a} \mid$ est le nombre de produits amiantés dans les diagnostics, et $\mid p_{diag} \mid$ est le nombre total des produits dans les diagnostics.

En utilisant l'équation 5.2, nous trouvons :

$$\alpha = \frac{\sum_{p_j \in F(p_k)} p_{ext}(p_j, d)}{\mid p_j \mid \times p_a(p_k, d) \times (1 + \omega)} - 1 \Rightarrow \alpha = \frac{-\omega}{1 + \omega} \tag{5.5}$$

La valeur de la proportion α et du réajustement w varient en fonction de l'année et de la classe de produit. Cependant, dans le cas où seul un petit nombre de diagnostics sont disponibles, nous pouvons calculer le réajustement une fois pour toutes les classes et toutes les années, et donc pour toutes les instances de la classe "Product" avec la formule suivante :

$$\omega = \frac{\mid p_{diag,a} \mid}{\mid p_{diag} \mid} - \frac{\sum_{p_j \in Product} p_a(p_j)}{\mid p_j \mid}$$

Une probabilité calculée p'_a est associée à chaque produit utilisé dans un bâtiment déjà décrit dans le graphe de connaissance du CSTB via la propriété "has_Probability" du concept "Calculated characteristic".

Pour modéliser les probabilités amiante calculées (Figure 4.5) dans l'ontologie Amiante, nous avons défini le concept de caractéristique calculée "Calculated characteristic" qui décrit les caractéristiques Amiante supposées du produit inconnu qui a été utilisé dans un bâtiment. Une instance de ce concept est décrite par la probabilité calculée et la classe de la probabilité. Seulement deux classes de probabilité ont été définies : forte et faible qui vont être représentées par 1 et 0 dans la propriété "has_Class". Pour classifier une probabilité comme forte ou faible, nous la comparons avec un seuil à déterminer expérimentalement en utilisant un ensemble de diagnostics pour trouver le meilleur seuil qui maximise l'*accuracy*. Si la probabilité est inférieure ou égale au seuil, alors elle sera considérée comme une probabilité faible, et elle sera forte dans le cas contraire.

5.5 Probabilités de présence d'amiante dans les parties du bâtiment

Pour calculer la probabilité de présence d'amiante dans une localisation, une structure ou un bâtiment, nous avons défini, en accord avec l'expert, une stratégie pessimiste qui propage la valeur maximum de l'ensemble des valeurs de probabilité de présence d'amiante dans les produits.

Ainsi, pour une localisation l_i, nous considérons tous les produits p_k qui composent la localisation l_i :

$$p_a(l_i) = Max(p_a(p_k))$$

De même, pour les structures s_i, il s'agira du maximum des probabilités de ses localisations l_k :

$$p_a(s_i) = Max(p_a(l_k))$$

Finalement, pour les bâtiments b_i, nous choisissons la valeur maximale des probabilités de ses structures s_k :

$$p_a(b_i) = Max(p_a(s_k))$$

Pour sauvegarder ces probabilités ainsi que la classe de chaque probabilité, nous avons ajouté les propriétés "has_Probability" et "has_Class" aux concepts : "Location", "Structure" et "Building" pour associer chaque composant à une probabilité et une classe. Associer une probabilité et une classe à tout un bâtiment, toute une structure ou toute une localisation permet à l'expert de facilement sélectionner les bâtiments et les éléments qu'il faut prioriser dans un programme de repérage de l'amiante.

5.6 Règles de propagation d'une probabilité à de nouveaux produits

Nous avons défini des règles SWRL qui peuvent être exploitées par les raisonneurs pour générer des probabilités d'amiante pour les produits présents dans la description d'un bâtiment nouvellement saisi. Pour éviter de recalculer une probabilité pour un nouveau produit, l'idée est d'exploiter les probabilités déjà décrites pour des produits de même classe (colle, enduit, etc.). Les ressources montrent que si un produit commercialisé est amianté une année donnée, alors celui-ci est amianté pour toutes les années antérieures. De même s'il devient non amianté pour une année spécifiée, il restera non amianté pour toutes les années suivantes.

Aussi, pour générer des probabilités d'amiante manquantes pour les produits, nous définissons dans le modèle de règle R_1 (modèle de règle défini pour chaque type de produit T) que si un produit de type T contient de l'amiante pour une année donnée Y_1, alors tous les produits de même type T contiennent de l'amiante pour toutes les années précédentes $Y_2 \leq Y_1$.

Rule pattern R_1 : T(P1), T(P2), has_Calculated_Characteristic(P1, C1), has_Calculated_Characteristic(P2, C2), contain(L1, P1), has_Location(S1, L1), has_Structure(B1, S1), has_Year(B1, Y1), contain(L2, P2), has_Location(S2, L2), has_Structure(B2, S2), has_Year(B2, Y2), has_Class(C1,1), lessThanOrEqual(Y2, Y1) -> has_Class(C2,1)

De même, si un produit de type T ne contient pas d'amiante à une date Y_1, alors un modèle de règle R_2 instanciée permet de déduire que tous les produits de même type T ne contiennent pas d'amiante pour toutes les années suivantes $Y_2 \geq Y_1$.

Rule pattern R_2 : T(P1), T(P2), has_Calculated_Characteristic(P1, C1), has_Calculated_Characteristic(P2, C2), contain(L1, P1), has_Location(S1, L1), has_Structure(B1, S1), has_Year(B1, Y1), contain(L2, P2), has_Location(S2,

L2), has_Structure(B2, S2), has_Year(B2, Y2), has_Class(C1,0), greaterThanOrEqual(Y2, Y1) -> has_Class(C2,0)

L'utilisation de règles de raisonnement pour déduire la classe de probabilité (0 ou 1) d'un nouveau produit a permis d'optimiser cette approche en réduisant le nombre de calculs de probabilité (Formule 5.2) d'une part et de son réajustement (Formule 5.4) d'autre part. Sachant que nous avons besoin d'utiliser des requêtes SPARQL pour récupérer les données nécessaires pour les deux formules, c.-à-d. nous récupérons les produits commercialisés pour la Formule 5.2, en plus des diagnostics et des diagnostics amiantées pour la Formule 5.4. En effet le raisonnement avec les règles et plus rapide que le processus de calcul.

5.7 Expérimentations

Afin de tester cette approche hybride, nous avons utilisé le jeu de données de CSTB et les ressources externes pour enrichir et peupler l'ontologie ASBESTOS en conservant les données originales et en calculant les caractéristiques des données fusionnées. L'objectif est d'observer l'évolution de la probabilité extraite au fur et à mesure des années sur les résultats obtenus et de montrer comment ces probabilités ont été réajustées. Nous avons ensuite calculé la probabilité de présence d'amiante dans un bâtiment et dans les éléments le composant et nous avons évalué la qualité de cette approche en comparant les diagnostics des produits avec les résultats obtenus. Nous avons comparé les résultats obtenus par l'approche hybride avec ceux obtenus par deux autres approches d'apprentissage de règles AMIE3 [27] et TILDE [2] et une approche naïve qui utilise que l'année de construction du bâtiment pour décider si un produit est amianté ou non, cette comparaison est appliquée aux mêmes jeux de données. Enfin, nous avons vérifié que cette approche était applicable quand peu de diagnostics sont disponibles en comparant ces résultats avec les résultats obtenus quand un petit sous-ensemble de diagnostics est utilisé pour l'étape de réajustement.

Dans cette section nous décrivons tout d'abord le résultat de l'étape d'enrichissement et de peuplement de l'ontologie avec les données issues des deux ressources externes, puis nous présentons le jeu de données fourni par le CSTB. Enfin nous décrivons l'évaluation quantitative et qualitative des résultats obtenus.

5.7.1 Jeux de données fournis par le CSTB

Nous disposons de deux ensembles de diagnostics de CSTB. Le premier jeu de données contient 2998 produits diagnostiqués qui appartiennent à 341 localisations, 214 structures et 94 bâtiments dont l'année de construction varie de 1948 à 1996. Parmi ces produits 1525 sont amiantés tandis que 1473 ne sont pas amiantés. L'amiante est présent dans 108 localisations sur les 341 existantes, dans 70 structures sur les 214 structures existantes, et dans 42 bâtiments sur 94.

Le deuxième jeu de données décrit seulement 50 produits diagnostiqués, qui appartiennent à 31 localisations,

27 structures et 16 bâtiments. Parmi les 50 produits, 34 produits sont amiantés alors que 16 d'entre eux sont sans amiante. L'amiante est présent dans 19 localisations sur les 31 existantes, dans 17 structures sur les 27 structures existantes, et dans 15 bâtiments sur 16.

5.7.2 Enrichissement et Peuplement de l'ontologie ASBESTOS avec les jeux de données issues du CSTB

L'enrichissement de l'ontologie avec le premier jeu de données à ajouter 38 classes de produits tendis que l'ontologie du deuxième jeu de données contient 15 classes de produit.

La figure 5.3 montre un extrait des classes ajoutées à l'ontologie après l'enrichissement avec le premier jeu de données du CSTB.

5.7.3 Enrichissement et Peuplement des ontologies avec les données issues de l'INRS et de l'ANDEVA

Le fichier ANDEVA décrit 650 produits sous forme tabulaire. Chaque produit est décrit par :

— la classe de produit,

— un ensemble de noms de produits qui présentent des propriétés communes et le nom du fournisseur.

— la date à partir de laquelle les produits ne sont plus amiantés.

Le fichier INRS décrit 300 produits, sous forme d'un tableau qui contient :

— un ensemble de noms de produits,

— le nom du fournisseur.

— les intervalles de temps où les produits sont amiantés avec un degré d'incertitude.

— le type d'amiante.

— les types d'utilisation (c.-à-d. classes de produit).

Dans un premier temps, nous avons fusionné les données tabulaires de l'INRS et l'ANDEVA pour enrichir et peupler l'ontologie ASBESTOS qui va ainsi contenir toutes les informations de l'ANDEVA et l'INRS.

L'étape d'extraction des informations et de fusion des descriptions de produits commercialisés issus des deux ressources externes nous a permis de créer 64 sous-classes de produit (p. ex. enduit, colle, etc.) et 694 instances de produits commercialisés.

Toutes les classes de produit ne sont pas instanciées par les ressources externes. Cas des jeux de données 1 et 2.

Presque tous les produits de l'INRS sont également mentionnés dans l'ANDEVA $(256/300)$. Parmi ces produits, seulement 35 d'entre eux conduisent à des valeurs conflictuelles pour les caractéristiques amiantes. Certains noms

FIGURE 5.3 – Extrait de sous-classes de produits ajoutées à la première ontologie (premier jeu de données)

de produit ont évolué au court du temps et leur nouveau nom est mentionné en même temps que l'ancien dans l'INRS (c.-à-d. c'est le cas pour 10 produits pour lequel nous gardons les deux labels possibles).

Le graphe de la Figure 5.4 montre l'évolution de la probabilité issue de la fusion pour l'ensemble des produits commercialisés (i.e propriété $has\_Probability$). La probabilité de présence d'amiante reste globalement stable jusqu'à 1972, puis elle décroît jusqu'à atteindre 0 en 1997, année où l'amiante a été interdit. Ce graphe confirme que

si une classe de produit est amiantée à une année donnée, alors elle est amiantée pour toutes les années antérieures. De même si elle devient non amiantée à une année, elle restera non amiantée à partir de cette année. La proportion de produits amiantés varie de $92,7\%$ à $44,8\%$ et nous savons que cette probabilité est surestimée, car elle est calculée uniquement sur des produits commercialisés qui ont été amiantés ou qui sont suspects.

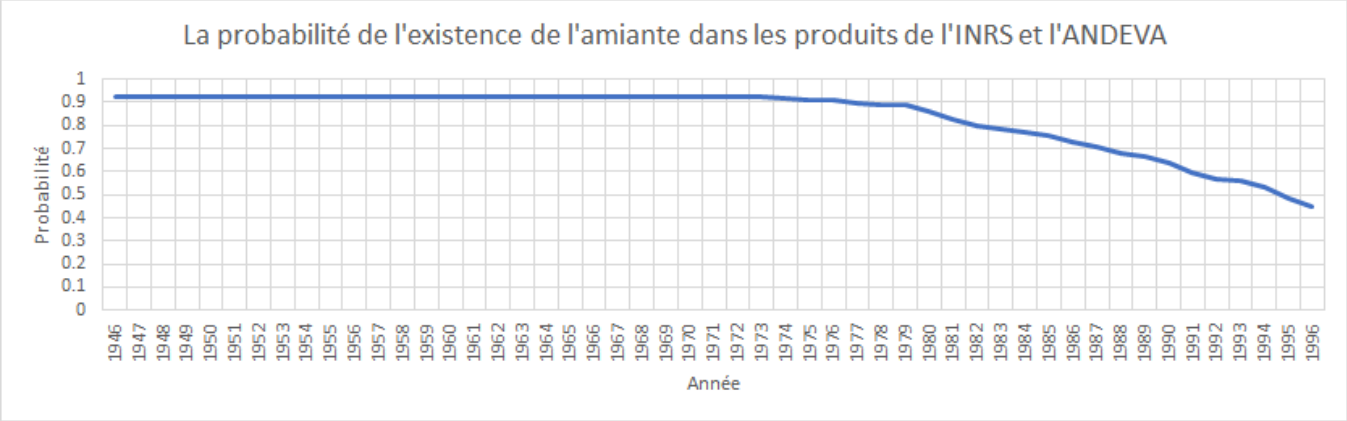


FIGURE 5.4 – La probabilité de l'existence d'amiante dans les produits de l'INRS et l'ANDEVA en fonction des années

5.7.4 Analyse quantitative des résultats

Calcul de la probabilité de présence d'amiante dans les éléments de bâtiment

Nous avons utilisé les diagnostics posés sur les 2998 produits du premier jeu de données pour calculer les valeurs du coefficient α qui permet de réajuster le nombre de produits commercialisés 5.2. Comme ce jeu de données comporte un grand nombre de produits, α est calculé selon deux facteurs, l'année de construction et le type de produit (voir Formule 5.5).

Les expérimentations montrent que les valeurs de α varient de 0 à $9,4$. Des variations existent pour une même classe de produit et des années différentes (p. ex. l'enduit en 1985 a un coefficient de réajustement, $\alpha = 0,37$ mais en 1986 $\alpha = 0,39$). De même, pour une même année, α varie selon la classe de produit (p. ex. en 1991, le coefficient de réajustement α pour l'enduit est $0,78$ mais pour le joint ce coefficient vaut $5,31$).

La Figure 5.5 montre sur les quatre classes de produits présentées dans le Tableau 5.4 que la probabilité de l'existence de l'amiante réajustée diffère d'une classe de produit à l'autre. Par exemple, les adhésifs amiantés sont peu nombreux et se sont désamiantés ou n'ont plus été commercialisés plus rapidement que les trois autres classes de produits présentées en exemple.

Quelle que soit la classe de produits considérée, la probabilité d'utilisation de produits amiantés diminue en fonction du temps. En effet, de moins en moins de produits commercialisés sont amiantés et en 1997, un produit est désamianté ou n'est plus commercialisé.

Famille de produits	Nombre de produits
Adhésif	5
Colles	31
Enduits	19
Mastics	61

TABLE 5.4 – Caractéristiques des classes de produits choisis pour le graphe 5.5

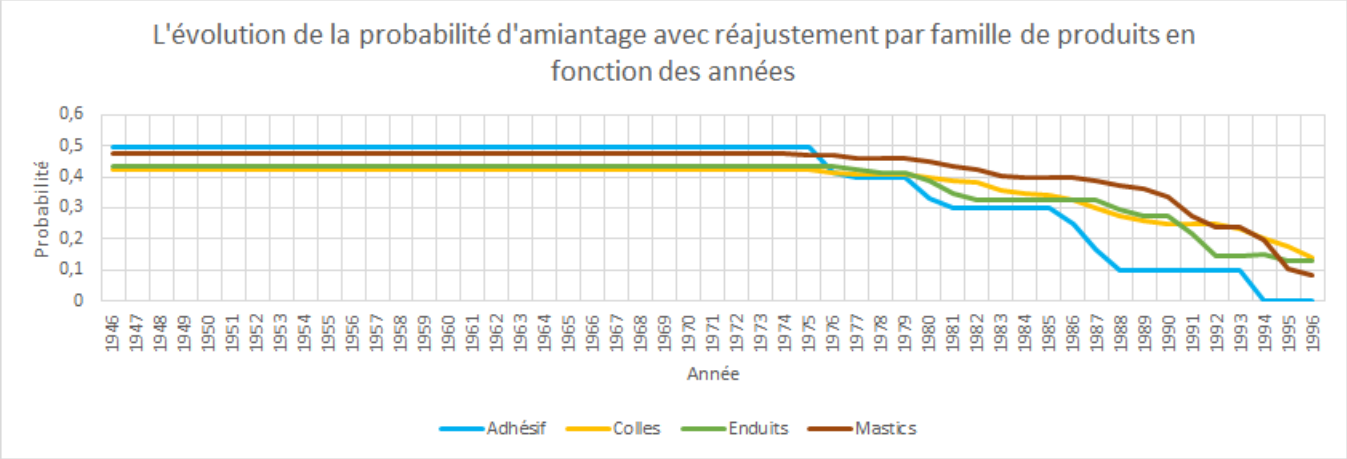


FIGURE 5.5 – L'évolution de la probabilité de présence d'amiante par classe de produits en fonction des années

Calcul du seuil

Nous avons ensuite calculé la probabilité de présence d'amiante pour les 2298 produits du premier jeu de données du CSTB et nous avons calculé le seuil pour classifier les produits selon leurs probabilités en produits amiantés (avec une forte probabilité notée par 1) ou non amiantés (avec une faible probabilité notée par 0).

Le seuil est déterminé de sorte qu'il maximise les vrais positifs et les vrais négatifs et minimise les faux positifs et les faux négatifs, c.-à-d. pour maximiser l'*accuracy*. Afin de déterminer le meilleur seuil nous avons utilisé la validation croisée sur trois tiers de données, c.-à-d. nous avons divisé les données dont nous disposons en trois tiers et à chaque fois nous effectuons l'apprentissage du seuil sur deux tiers et nous le testons sur le tiers restant.

La Figure 5.6 montre les différentes valeurs de la F-mesure moyenne et l'*accuracy* en fonction du seuil. Lorsque la valeur du seuil appartient à l'intervalle $[0,34, 0,45]$ nous obtenons les meilleurs F-mesure ($0,89$ en moyenne) et la meilleure *accuracy* ($0,97$ en moyenne). Nous avons donc choisi le seuil médian $s = 0,39$. Aussi pour classifier les produits, nous comparons leur probabilité avec $s = 0,39$, si la probabilité $p_a \leq s$ alors elle est faible et si $p_a > s$ alors elle est considérée comme forte.

5.7.5 Analyse qualitative des résultats

Nous avons tout d'abord évalué l'approche hybride sur l'échantillon disponible de diagnostics amiante qui décrivent les tests effectués sur les 2998 produits du premier jeu de données.

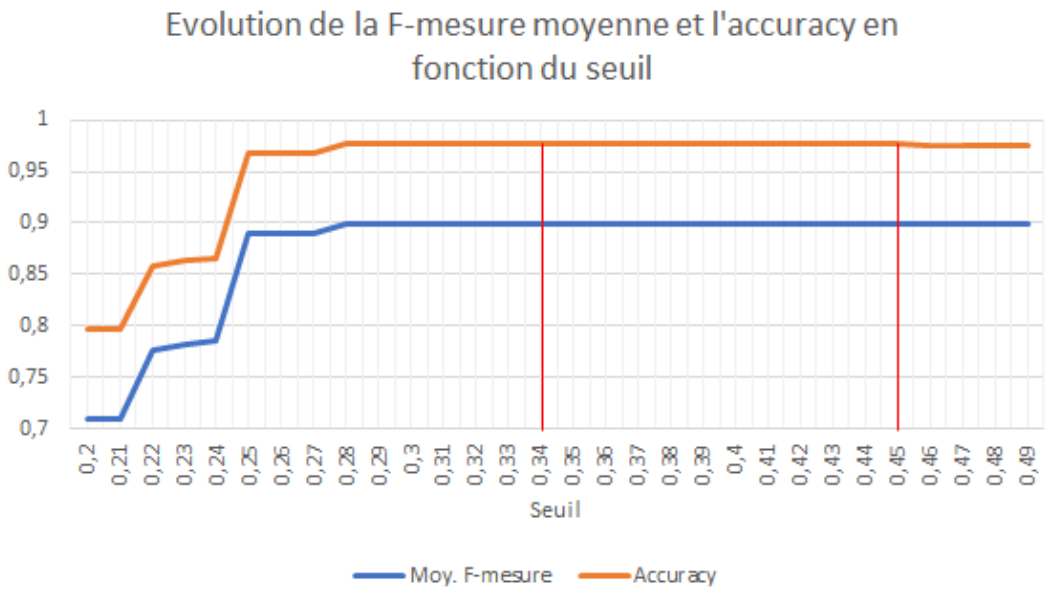


FIGURE 5.6 – L'évolution de la F-mesure moyenne et l'*accuracy* en fonction du seuil

Nous avons montré à l'expert du CSTB un échantillon constitué d'une vingtaine de produits avec leurs diagnostics prédits choisis aléatoirement. Les probabilités se répartissaient en deux classes "amiante" (faible probabilité et forte probabilité). La Table 5.5 montre un exemple de 3 des résultats expertisés. Pour chaque produit, une explication de la probabilité calculée peut être constituée de la probabilité elle-même, de l'extrait des ressources INRS et ANDEVA comportant les produits commercialisés de la même classe et de la même année, et la proportion des produits diagnostiqués de la même classe qui sont amiantés dans des bâtiments construits dans la même année. La Table 5.6 montre les différents éléments d'information qui peuvent être montrés pour le premier exemple de la Table 5.5. Lorsque nous utilisons les probabilités de la fusion de table 5.6 avec la Formule 5.1 nous trouvons la probabilité non réajustée $p_a(bande\_57, 1986) = 0,94$. En plus, nous avons 28 produits diagnostiqués de type "Bande", 14 produits entre eux sont diagnostiqués positifs. Nous calculons le ω avec l'équation 5.4 : $\omega = \frac{14}{28} - 0,94 = -0,44$. L'équation 5.5 donne $\alpha = 0,78$. Cette valeur de α va être utilisée pour calculer la probabilité réajustée de la Formule 5.2 : $p'_a(bande\_57, 1986) = 0.53$. Le seuil utilisé pour classer cette probabilité comme forte est celui appris dans les tests ($s = 0,39$). L'expert a validé le processus de calcul de la probabilité à partir des produits commercialisés sur l'ensemble des exemples.

Nous avons évalué les résultats en calculant la précision, le rappel et la F-Mesure pour les positifs et les négatifs, ainsi que l'*accuracy* et la couverture qui mesure la proportion des produits pour lesquels une classe peut être prédite dans l'ensemble des produits. La Table 5.7 montre l'ensemble des résultats et utilise les notations suivantes :

— TP (vrais positif) : représente un produit positif est classé comme étant un positif.

— TN (vrais négatif) : représente un négatif positif est classé comme étant un négatif.

— FP (faux positif) : représente un produit négatif est classé comme étant un positif.

Code du produit utilisé	Type de produit	Année	ID du bâtiment	Probabilité	Classe de probabilité
bande_57	Bande	1986	Bat_8	0,53	1 (forte)
collesetjointsdecarrelageragreages primairesdaccrochage_2951	Colles et joints de carrelage ragréage primaire d'accrochage	1992	Bat_89	0,25	0 (faible)
enduitsderagreage debullagelissage_62	enduits de ragréage de bullage lissage	1986	Bat_8	0,56	1 (forte)

TABLE 5.5 – Échantillon de résultats expertisés

Produits commercialisés	Probabilité de l'INRS en 1986	Probabilité de l'ANDEVA en 1986	Probabilité de la fusion en 1986
BANDE 703	1	1	1
BANDE 709	1	1	1
BANDE 714 E	1	1	1
BANDE 733/734	1	1	1
BANDE 731/732	1	1	1
BANDE 737 à 739	1	1	1
BANDE 747 à 749	1	1	1
DIATISOL	0,5	0,5	0,5

TABLE 5.6 – Les probabilités extraites et fusionnées des produits commercialisés de type “Bande” en 1986

— FN (faux négatif) : représente un produit positif est classé comme étant un négatif.

— UP (indécidable positif) : représente un produit positif non classé.

— UF (indécidable négatif) : représente un produit négatif non classé.

La table 5.7 montre que l'approche hybride obtient en moyenne une très bonne précision que ce soit pour les positifs (97%) ou les négatifs (99%) ainsi qu'une bonne *accuracy* ($0,97$). En revanche, l'approche ne permet pas de classifier l'ensemble des produits (c.-à-d. couverture de 83%). En effet, nous avons vu que certaines classes ne sont pas représentées dans les ressources externes. Or, il est nécessaire, pour calculer la probabilité de présence d'amiante, qu'il existe un produit commercialisé de la même classe la même année. Cependant, ces résultats signifient que si l'on considère 1000 produits, théoriquement, des prélèvements pour détecter la présence d'amiante devraient être effectués pour chacun d'entre eux. Si l'expert du CSTB ne diagnostic les produits détectés comme positifs auxquels on peut ajouter les produits non classés, il n'effectuera que 646 prélèvements ($64,6\%$), et il n'y aura que 5 produits amiantés non testés ($0,98\%$ des produits amiantés).

Nous avons tout d'abord effectué une comparaison avec deux *baselines* : une première *baseline* qui se base uniquement sur l'année de construction du bâtiment pour décider qu'un produit est amianté ou non et une deuxième *baseline* qui se base uniquement sur la classe de produits. Pour la première *baseline*, la meilleure année seuil a été calculée à partir de l'ensemble des diagnostics du jeu de données. De même, pour la deuxième *baseline*, la décision suit la classe majoritaire dans la classe de produit concernée. Nous avons également comparé les résultats avec deux autres approches d'apprentissage de règles : AMIE3 [27] et TILDE [2] (voir la Table 5.7). Il s'agit en effet d'approches qui permettent d'effectuer des classifications qui sont de nature différente. AMIE3 permet

de rechercher l'ensemble de toutes les règles connexes et fermées qui ont un *head coverage* et une confiance supérieurs à un seuil spécifié, dans un graphe de connaissance. Tilde découvre des arbres de décision dans des données relationnelles et permet de couvrir par construction toutes les instances des classes considérées.

Nous avons paramétré AMIE3 pour qu'il ne considère que les règles qui concluent sur les prédicats "has_Classe (CalculatedCharacteristic, 1)" ou "has_Classe(CalculatedCharacteristic, 0)". AMIE3 cherche par défaut des règles d'une longueur de 3 atomes au total (c.-à-d. prémisse et conclusion). Nous avons augmenté cette longueur à 4 et 6 pour améliorer la précision de ses résultats. Le *head coverage* a été fixé à $0,001$ et la confiance à $0,6$ afin d'obtenir une bonne couverture des données.

Afin de trouver la meilleure année seuil pour la première *baseline* nous avons suivi l'évolution de l'*accuracy* et de la F-mesure en fonction de l'année et nous avons trouvé que la meilleure année seuil pour la *baseline* 1 est 1986 (Figure 5.7).

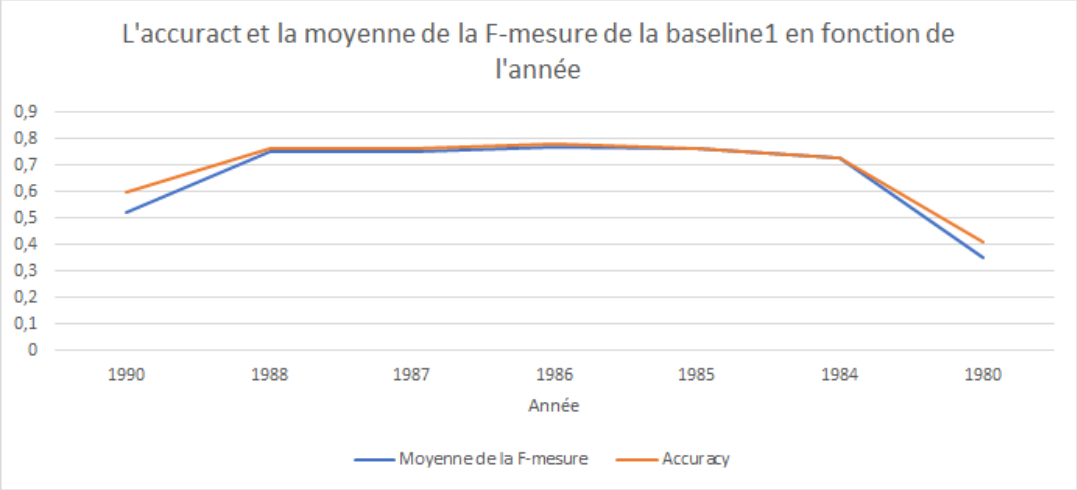


FIGURE 5.7 – L'évolution de la F-mesure moyenne et l'*accuracy* de la *baseline* 1 en fonction de l'année

La table 5.7 montre que l'approche hybride obtient une précision significativement plus grande que celle de la *baseline* 1 pour les positifs (97% contre 71% pour la *baseline* 1) ainsi que les négatifs (99% contre 95% obtenu par la *baseline* 1). Ce qui a guidé vers une plus grande *accuracy* pour l'approche hybride $0,97$ comparant à $0,78$ pour la *baseline* 1. La couverture de 100% obtenu par la première *baseline* est expliquée par la stratégie suivie, il suffit qu'un produit soit associé à un bâtiment avec une année pour pouvoir obtenir une décision ce qui est le cas pour tous les produits.

La *baseline* 2 a également obtenu des précisions plus petites que l'approche hybride pour les positifs (97% contre 75% pour la *baseline* 2) ainsi que les négatifs (99% contre 77% obtenu par la baselline 2). Aussi une plus faible *accuracy* $0,76$. La table 5.7 montre que la *baseline* 1 a obtenue une meilleure *accuracy* que la deuxième *baseline* ($0,78$ contre $0,76$) ainssi qu'une meilleure précision pour les négatifs (95% contre 77%) avec une précision des positifs moins bonne, mais proche de celle obtenue par la *baseline* 2 (71% pour la *baseline* 1 et 75% pour la

Système	L'approche hybride	AMIE3 $l = 4$	AMIE3 $l = 6$	TILDE	*Baseline* 1 année de construction	*Baseline* 2 type de produit
TP	465	381	473	431	507	419
TN	348	288	264	358	271	339
FP	16	146	226	128	207	139
FN	5	74	32	87	14	102
UP	38	54	3	0	0	0
UN	127	58	0	0	0	0
Pos. précision	**97%**	72%	68%	77%	71%	75%
Pos. rappel	92%	75%	93%	83%	**97%**	80%
Pos. F-mesure	**0,94**	0,73	0,79	0,80	0,82	0,77
Neg. précision	**99%**	80%	89%	80%	95%	77%
Neg. rappel	**71%**	59%	54%	74%	57%	**71%**
Neg. F-mesure	**0,83**	0,68	0,67	0,77	0,71	0,74
Moy. F-mesure	**0,89**	0,71	0,73	0,79	0,77	0,76
Accuracy	**0,97**	0,75	0,74	0,79	0,78	0,76
Couverture	83%	89%	**100%**	**100%**	**100%**	**100%**

TABLE 5.7 – Comparaison entre l'approche hybride, AMIE3 avec $l = 4$ et $l = 6$ (minHC=0,001, minConf=0,6), TILDE et les deux *baselines*

baseline 2).

Nous remarquons dans cette table aussi que l'approche hybride obtient une meilleure précision pour les positifs et pour les négatifs (97% et 99% respectivement) comparée à AMIE3 (72% pour les positifs lorsque $l = 4$ et 89% pour les négatifs lorsque $l = 6$) et TILDE (77% pour les positifs et 80% pour les négatifs). L'approche hybride obtient aussi une meilleure F-mesure ($0,94$ pour les positifs et $0,83$ pour les négatifs) comparée à AMIE3 ($0,79$ pour les positifs et $l = 6$ et $0,68$ pour les négatifs lorsque $l = 4$) et TILDE ($0,80$ pour les positifs et $0,77$ pour les négatifs). L'approche hybride obtient la meilleure *accuracy* de $0,97$ ($0,75$ pour AMIE3 avec $l = 4$ et $0,79$ pour TILDE). Cette *accuracy* très haute s'explique par le petit nombre de fausses décisions obtenues par l'approche hybride (16 FP et seulement 5 FN qui représentent respectivement 3% de décisions positives et 1% de décisions négatives). Le fait d'exploiter les informations issues des ressources externes permet à l'approche hybride de prendre de meilleures décisions. En revanche, il faut noter que les deux approches concurrentes peuvent exploiter toute la description du produit (p. ex. le type de localisation qui le contient, la structure, etc.), tandis que notre approche ne se base que sur la classe du produit et l'année de construction.

À cause de l'incomplétude des données externes utilisées par l'approche hybride, celle-ci a obtenu une couverture moins bonne que celle obtenue par les deux autres approches (83%). Cette différence de couverture de données revient au fait que les données externes sont incomplètes et elles ne décrivent que les produits qui ont été soupçonnés d'avoir de l'amiante durant au moins une période de commercialisation, c.-à-d. les données externes ne décrivent pas les produits qui n'ont jamais été amiantés. Nous pouvons observer ce fait, car le nombre de positifs non-classés (UP = 38) est plus faible que le nombre de négatifs non-classés (UN = 127).

Optimisation du processus de classification avec l'utilisation des règles de propagation

Nous avons testé et comparé le calcul de probabilité et son réajustement avec la propagation des probabilités avec les règles de raisonnement. Cette comparaison a été effectuée sur le premier jeu de données. L'approche hybride a pris $2h10min$ pour calculer et réajuster la probabilité pour l'ensemble de test, alors qu'elle n'a pris que $2sec$ avec la propagation de probabilité avec l'application des règles de raisonnement par Pellet sur le même jeu de données. A cause de cette différence dans le temps d'exécution des deux méthodes, nous avons privilégié l'utilisation des règles de propagation afin d'optimiser le nombre de calculs, et donc le temps de calcul.

Evaluation de l'approche hybride quand seulement un petit nombre de diagnostics est utilisé

En utilisant les ressources externes (ANDEVA et INRS), nous avons calculé la probabilité de l'existence de l'amiante pour chaque classe de produits et pour chaque année (de 1946 jusqu'à 1997) en utilisant seulement le petit nombre de diagnostics décrits dans le deuxième jeu de données du CSTB (50 produits diagnostiqués). L'objectif est de savoir si l'approche hybride peut fonctionner même avec un petit nombre de diagnostics.

Comme nous possédons peu de diagnostics dans ce deuxième test, nous avons calculé une seule valeur de coefficient d'ajustement $\alpha \approx 1,0076$ pour tous les produits et pour toutes les années. Nous avons ensuite utilisé la méthode de test "Leave-one-out" pour déterminer le meilleur seuil, c.-à-d. à chaque test nous apprenons le seuil avec 49 produits et nous le testons avec le produit restant. Nous avons constaté que le meilleur seuil est le même pour tous les sous-ensembles possibles de 50 produits ($s = 0,25$). Enfin, nous avons utilisé ce seuil pour classifier les produits, les localisations, les structures et les bâtiments qui ont une forte probabilité de contenir de l'amiante.

Le graphe de la Figure 5.8 compare la probabilité sans réajustement et avec ce réajustement uniforme. Comme nous l'avons vu, sans le réajustement, la proportion de produits amiantés varie de $92,7\%$ à $44,8\%$. Quand ce réajustement est appliqué, la proportion maximum de produits amiantés devient $46,17\%$.

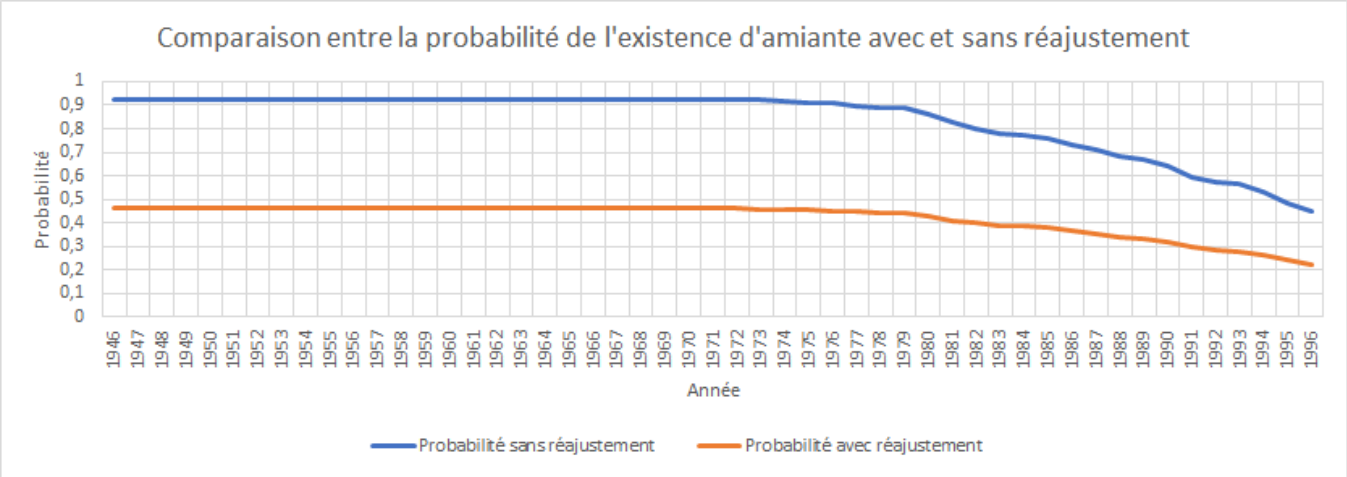


FIGURE 5.8 – Comparaison entre la probabilité de l'existence d'amiante dans les produits de l'INRS et l'ANDEVA avec et sans réajustement

Produits		Localisations		Structures		Bâtiments	
TP=34	FP=9	TP=19	FP=6	TP=17	FP=6	TP=15	FP=0
FN=0	TN=7	FN=0	TN=6	FN=0	TN=4	FN=0	TN=1
UP=0	UN=0	UP=0	UN=0	UP=0	UN=0	UP=0	UN=0

TABLE 5.8 – Matrice de confusion pour les produits, les emplacements, les structures et les bâtiments

	Moy. F-mesure	*Accuracy*
Produits	0.75	0.82
Localisations	0.77	0.81
Structures	0.71	0.78
Bâtiments	1	1

TABLE 5.9 – La F-mesure et l'*accuracy* pour tous les composants du bâtiment (s=0,25)

Le tableau 5.8 représente la matrice de confusion obtenue pour toutes les parties du bâtiment. Le tableau 5.9 montre que nous pouvons classer correctement plus de 80% des produits en utilisant le seuil choisi (*accuracy* de $0,82$). De plus, notre approche pessimiste nous permet d'obtenir une *accuracy* de $0,81$ (respectivement $0,78$) pour les localisations (respectivement les structures) et une *accuracy* de 1 pour les bâtiments. Cela signifie que tous les bâtiments qui ont une forte probabilité d'amiante ont au moins un composant qui contient de l'amiante alors que tous les bâtiments à faible probabilité ne contiennent aucun produit amianté.

La table 5.10 montre que lorsque nous disposons plus grand nombre de diagnostics, nous obtenons des résultats plus précis pour les positifs (97% pour le premier jeu de données et 79% pour le deuxième) et elle est légèrement moins précis vers 99% pour les négatifs. La disponibilité de plusieurs produits diagnostiqués améliore considérablement l'*accuracy* pour obtenir $0,97$ contre $0,82$ lorsque seulement un petit nombre de diagnostics est disponible dans le premier jeu de données.

5.8 Conclusion

Dans ce chapitre, nous avons présenté la première approche de prédiction de la présence d'amiante dans un bâtiment qui combine des méthodes statistiques et des méthodes basées sur des règles pour prédire la présence d'amiante dans un bâtiment. Cette approche a pour objectif d'aider un opérateur de repérage à prioriser les bâtiments dans lesquels il faut effectuer un prélèvement et à prioriser les prélèvements à effectuer dans ce bâtiment. L'approche se base sur l'hypothèse que même si le produit commercialisé utilisé dans le bâtiment n'est pas connu, il est possible d'exploiter des ressources externes qui listent des ensembles de produits existants sur le marché au moment de la construction du bâtiment et qui précise si ces produits sont amiantés.

Pour cela, nous avons tout d'abord ajouté à l'ontologie ASBESTOS un module qui permet de représenter les données des ressources externes ainsi que les probabilités calculées et les classes prédites pour chaque produit utilisé dans un bâtiment. Ce module permet de représenter des données temporelles probabilistes sur la présence

Approche hybride	Premier jeu de données	deuxième jeu de donnée
TP	465	34
TN	348	7
FP	16	9
FN	5	0
UP	38	0
UN	127	0
Pos. précision	**97%**	79%
Pos. rappel	92%	**100%**
Pos. F-mesure	**0,94**	0,88
Neg. précision	99%	**100%**
Neg. rappel	**71%**	44%
Neg. F-mesure	**0,83**	0,61
Moy. F-mesure	**0,89**	0,75
Accuracy	**0,97**	0,82
Couverture	83%	**100%**

TABLE 5.10 – Comparaison entre les résultats de l'approche hybride avec le premier et le deuxième jeu de données

d'amiante dans les produits commercialisés décrits dans les ressources externes dont on garde la provenance. Nous avons ensuite proposé une méthode pessimiste de calcul des probabilités de présence d'amiante qui se base sur ces données incertaines et incomplètes et sur un ensemble de diagnostics.

Pour faire face à l'incertitude, nous avons fusionné les données conflictuelles en appliquant une méthode pessimiste pour limiter les risques. Pour résoudre l'incomplétude due à l'absence de certains produits commercialisés négatifs dans les ressources externes, nous avons réajusté les probabilités calculées avec un échantillon de diagnostics qui contient des descriptions de produits positifs et négatifs.

Les résultats des expérimentations montrent que l'utilisation des ressources externe permet d'obtenir une *accuracy* très élevée et que les résultats sont significativement meilleurs que ceux obtenus par d'autres méthodes. De plus, comme l'approche hybride se base essentiellement sur ces ressources externes, elle peut retourner de bons résultats même avec un petit nombre de diagnostics. Son utilité pour prioriser des prélèvements a pu être attestée par ces expérimentations, car très peu de produits amiantés ne seraient pas testés si l'expert considère uniquement les produits classés comme ayant une forte probabilité de présence d'amiante ainsi que les produits non classés. En revanche, compte tenu de l'incomplétude des ressources externes, l'approche proposée ne permet pas de prendre une décision pour tous les produits. Une autre approche est nécessaire pour utiliser la richesse des descriptions des produits dans les diagnostics, et en particulier le contexte de leur utilisation dans un bâtiment qui peut influencer la présence d'amiante.

Dans le chapitre suivant, nous allons montrer l'utilité du contexte dans lequel un produit a été utilisé et son influence sur la probabilité de présence d'amiante. Plus précisément, nous allons utiliser les caractéristiques des bâtiments, des structures, des localisations et des produits qui peuvent influencer sur la présence d'amiante dans les éléments de bâtiments, en utilisant la sémantique de l'ontologie.

Chapitre 6

Découverte de règles contextuelles de prédictions à partir de données de diagnostics amiante

6.1 Introduction

Dans ce chapitre, nous présentons notre deuxième approche, basée sur l'ontologie ASBESTOS, qui découvre des règles qui peuvent être utilisées pour estimer la probabilité de l'existence de produits amiantés dans un bâtiment. L'approche proposée s'inspire des techniques de Programmation Logique Inductive (PLI) de type *générer et tester*, mais se concentre sur la découverte de règles qui décrivent le produit et son contexte par un ensemble de prédicats déclarés comme potentiellement pertinents par l'expert. Sur la base des relations de subsomption et des connaissances générales sur l'évolution de l'utilisation de l'amiante au fil des années, l'algorithme découvre un ensemble de règles qui prédisent la présence d'amiante dans les produits d'un composant de bâtiment. L'originalité de l'approche CRA-Miner [34, 35] est de se baser sur un contexte sémantique, des heuristiques dédiées aux propriétés de type partie-tout (*part-of*) omniprésentes dans les descriptions des bâtiments et des contraintes temporelles utilisant des seuils calculés.

Nous présentons dans la Section 6.2 notre approche CRA-Miner qui découvre des règles contextuelles pour prédire la présence d'amiante dans les produits d'un bâtiment. Nous discutons ensuite dans la Section 6.3 la complexité de l'espace de recherche de chacune des étapes de CRA-Miner. Nous présentons dans la Section 6.4 les résultats des expérimentations de CRA-Miner sur le jeu de données de CSTB et une comparaison avec d'autres approches de découverte de règles et avec l'approche hybride présentée dans le chapitre précédent. Nous conclurons ce chapitre avec la combinaison entre CRA-Miner et l'approche hybride, la stratégie suivie dans la

combinaison et les résultats obtenus (Section 6.5).

6.2 L'approche CRA-Miner

Dans cette section, nous décrivons tout d'abord les règles logiques contextuelles que nous voulons fournir aux experts pour les aider à détecter les matériaux contenant de l'amiante dans le bâtiment. Nous présentons ensuite l'algorithme CRA-Miner qui permet de générer ces règles à partir de l'ontologie ASBESTOS peuplée.

6.2.1 Règles contextuelles pour la prédiction de l'amiante

Une règle contextuelle pour la prédiction de l'amiante (CRA) est une conjonction de prédicats qui conclut sur la présence ou l'absence d'amiante dans un produit P. Nous considérons la borne supérieure hors-contexte de l'espace de recherche $\top$ suivante : $product(P), has\_diagnostic\_characteristic(P, D) \rightarrow has\_diagnostic(D, Value)$. L'ensemble des règles contextuelles qui peuvent être construites à partir de cette borne supérieure est défini en utilisant un contexte conceptuel. Ce contexte est utilisé par les experts pour sélectionner les éléments de l'ontologie décrivant le produit P pouvant avoir un impact sur la présence d'amiante. Ces prédicats utilisés pour spécialiser la règle représentent un biais de langage tel que défini en Programmation Logique Inductive (PLI) [39].

Définition 6 (Contexte conceptuel) *Un contexte conceptuel CO est défini par un sous-graphe de l'ontologie, c'est-à-dire un ensemble de classes et de propriétés, qui déterminera les prédicats utilisables dans le corps de la règle.*

Exemple 1 $CO = \{product, location, structure, contain, has\_location, has\_region, has\_year, has\_structure, has\_diagnostic\_characteristic\}$ est un exemple de contexte conceptuel.

Une règle contextuelle est basée sur le vocabulaire de l'ontologie sélectionné dans le contexte conceptuel et les spécialisations du prédicat $SWRL : CompareTo$ qui peut être ajouté pour introduire des contraintes sur l'année de construction du bâtiment (c.-à-d. intervalles ouverts) :

Définition 7 (Règle contextuelle) *Soit CO un contexte conceptuel, une règle contextuelle $\overrightarrow{B} \rightarrow h$, où $\overrightarrow{B} = \{B_1, B_2, ..., B_n\}$, est telle que $\forall B_i \in \overrightarrow{B}, \exists B_j \in CO \cup \{SWRL : CompareTo\}$ tel que $B_i \sqsubseteq B_j$ et h est le prédicat $has\_diagnostic$ qui est instancié par la valeur "positif" ou "négatif".*

Une règle contextuelle doit également respecter les propriétés de fermeture et de connectivité définies dans les approches de fouille de règles telles que [27].

Exemple 2 La règle suivante est une règle contextuelle connexe et fermée qui peut être formée avec le contexte CO défini dans l'exemple 1 :

colle(P), contain(L, P), has_location(S, L), peinture(P2), contain(L, P2), has_structure(B, S), has_year(B, Y), has_region(B, "Paris"), lessThanOrEqual(Y, "1950"), has_diagnostic_characteristic(P, D) $\rightarrow$ *has_diagnostic(D, "positif")*

Cette règle exprime qu'une colle présente dans un bâtiment parisien construit avant 1950, qui est utilisée dans la même localisation qu'une peinture, est potentiellement amiantée.

Des contraintes supplémentaires sont définies pour réduire la complexité du contexte et limiter la taille de l'espace de recherche pour les propriétés multi-valuées décrivant les parties de bâtiments (c.-à-d. $contain$, $has\_location$, et $has\_structure$).

L'expert peut tout d'abord définir le nombre maximum d'occurrences des autres composants du bâtiment qui peuvent apparaître dans le corps de la règle : $maxSibS$ est utilisé pour définir le nombre de structures frères de la structure qui contient le produit P, $maxSibL$ est le nombre maximum de localisations frères, et $maxSibP$ représente le nombre maximum de produits frères.

Example 3 Si l'expert considère que le type des autres structures présentes dans le bâtiment ne peut pas influencer la présence d'amiante dans P, alors $maxSibS = 0$ et l'approche ne pourra pas construire la règle suivante :

enduit(P), contain(L, P), has_location(S1, L), separateur_vertical(S1), has_structure(B, S1), has_structure(B, S2), sol(S2) , has_year(B, Y), has_region(B, "Lyon"), SWRL :lessThanOrEqual(Y, "1963"), has_diagnostic_characteristic(P, D) $\rightarrow$ *has_diagnostic(D, "positif")*

En effet, la structure $S2$ ne devrait pas être considérée (frère de $S1$ par la propriété $has\_structure$, $S1$ contenant le produit cible).

Enfin, les experts du CSTB considèrent que seule la prise en compte des types de produits les plus spécifiques peut impacter le choix du produit cible commercialisé utilisé et donc la présence d'amiante. Par exemple, la présence d'un revêtement dans la même localisation qu'un produit cible de type colle n'est pas significative tandis que la présence d'un revêtement de sol peut impacter le choix de la colle commercialisée utilisée. Une hypothèse similaire est réalisée pour les localisations et les structures. Aussi, seules les classes les plus spécifiques sont ajoutées dans les relations de type *part-of* considérées.

Pour mesurer la qualité des règles, nous utilisons les mesures de qualité classiques de *head coverage* (hc) [27] et de confiance ($conf$) qui ont été définies pour les règles relationnelles.

Le *head coverage* (hc) représente la ratio entre le support, c.-à-d. le nombre de prédictions correctes de $has\_diagnostic(D, \text{“positif”})$ (respectivement $has\_diagnostic(D, \text{“negatif”})$) générées par la règle, et le nombre de diagnostics $has\_diagnostic(D, \text{“positif”})$ (respectivement $has\_diagnostic(D, \text{“negatif”})$) qui sont présents dans le graphe de connaissance :

$$hc(\overrightarrow{B} \rightarrow has\_diagnostic(D, val)) = \frac{supp(\overrightarrow{B} \rightarrow has\_diagnostic(D,val)}{\#(D,val):has\_diagnostic(D,val)}$$

La confiance ($conf$) est définie par le ratio entre le support de la règle et le nombre de diagnostics différents qui participent à une instanciation du corps de la règle.

$$conf(\overrightarrow{B} \rightarrow has\_diagnostic(D, val)) = \frac{supp(\overrightarrow{B} \rightarrow has\_diagnostic(D,val))}{\#D:\exists X_1,...,X_n:\overrightarrow{B}}$$

L'objectif est de découvrir toutes les règles les plus générales qui sont conformes aux contraintes du biais de langage et qui ont $hc \geq minHc$ et $conf \geq minConf$.

6.2.2 Évolution de la présence d'amiante au fil du temps

Il a été démontré dans la Figure 5.8 dans les expérimentations du chapitre précédent que le nombre de produits commercialisés amiantés sur le marché reste stable jusqu'en 1972, puis diminue pour atteindre 0 en 1997, lorsque l'usage de l'amiante est devenu interdit en France. En effet, soit les produits ont été désamiantés, soit ils ont été abandonnés. Ainsi, même si la probabilité d'amiante diffère d'une classe de produit à une autre (p. ex., les adhésifs ont perdu leur amiante plus tôt et plus rapidement que d'autres classes de produits), nous savons que cette probabilité diminue avec le temps. Aussi, si une règle contextuelle conclut sur l'absence d'amiante pour les produits utilisés dans les bâtiments construits après une année donnée Y_1, la confiance ne pourra qu'être égale ou supérieure pour $Y_2 \geq Y_1$. Cette caractéristique est exploitée pour élaguer l'espace de recherche lorsque le prédicat $greaterThanOrEqual$ ou $lessThanOrEqual$ est généralisé.

6.2.3 Module de CRA-Miner dans l'ontologie

Le but de CRA-Miner est de classifier les données de CSTB (les produits) en deux classes : positif (amianté) et négatif (non amianté). Afin de lui permettre de sauvegarder ses résultats, nous avons ajouté à l'ontologie de base le module représenté par la Figure 6.1. Le module de l'ontologie de CRA-Miner contint un nouveau concept “Predicted characteristic” qui peut contenir la classe prédite pour un produit (la classe peut avoir la valeur de 1 dans le cas de positif et de 0 dans le cas de négatif).

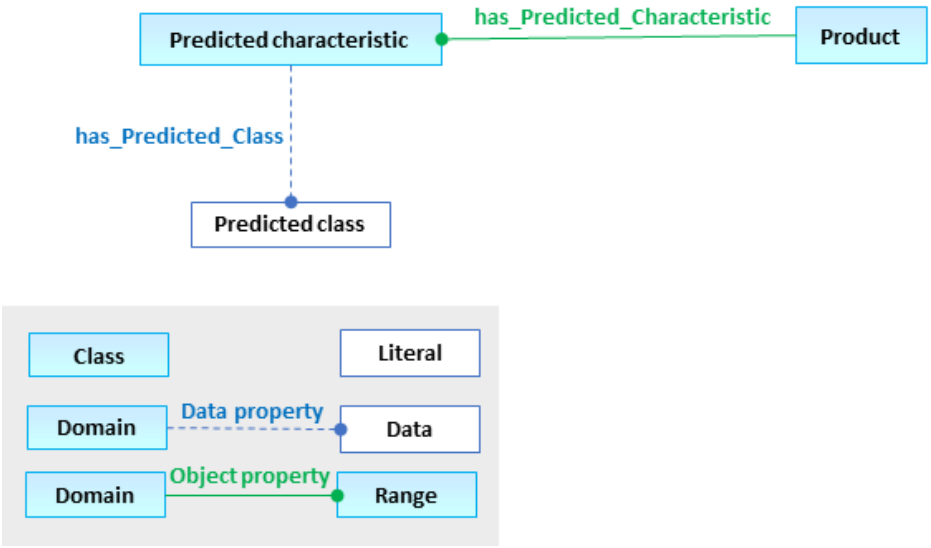


FIGURE 6.1 – Les concepts du module de CRA-Miner

6.2.4 Algorithme CRA-Miner

Le but de l'algorithme CRA-Miner est de générer toutes les règles contextuelles permettant de prédire la présence d'amiante dans les produits à partir des exemples positifs et négatifs décrits dans le graphe de connaissances (GC) et telles que $hc \geq minHC$ et $conf \geq minConf$. Ces règles seront utilisées pour prédire la présence d'amiante dans les produits en fonction de l'année, de la composition du bâtiment dans lequel il apparaît et de la région où il se situe.

L'algorithme de type *descendant générer et tester*, spécialise la borne supérieure de l'espace de recherche T en considérant la hiérarchie des classes de produit, en ajoutant des contraintes sur la localisation et la structure de ce produit, sur la présence de produits, localisations ou structures apparaissant dans le même composant, ainsi que des contraintes temporelles sur l'année de construction.

L'algorithme a comme entrées le graphe de connaissances, le biais de langage, un seuil $minConf$ sur la confiance, un seuil $minHC$ sur le *head coverage* de la règle, ainsi que les valeurs de $maxSibP$, $maxSibL$ et $maxSibS$ qui limitent le nombre de frères de produits, de localisations et de structures à ajouter à la règle. Le résultat est un ensemble $\mathcal{CR}$ de règles contextuelles.

L'exploration de l'espace de recherche est guidée par les relations de subsomption de l'ontologie (exploration *top-down* des produits cibles, de leurs localisations et de leurs structures) et exploite le fait que le nombre de produits amiantés est décroissant au fur et à mesure des années. A chaque étape de spécialisation, les règles construites qui possèdent dans une valeur de confiance et une valeur de *head coverage* plus grande que les seuils sont stockées dans l'ensemble $\mathcal{CR}$. Pour toutes les règles telles que $conf = 1$ ou $hc < minHc$, la spécialisation s'arrête.

Nous décrivons les étapes de l'algorithme pour le contexte le plus général qui a été défini par les experts du

CSTB, c.-à-d. le contexte CO défini dans l'exemple 1. L'algorithme comporte les 5 étapes suivantes :

1- **Spécialisation de $\top$ en utilisant des sous-classes de produit** :

Dans cette phase, nous remplaçons dans le $\top$ la classe $product$ par toutes les classes plus spécifiques (p. ex. enduit, peinture, etc.) tant que $hc \geq minHC$ et générons donc toutes les règles "*hors-contexte*" qui peuvent être trouvées pour chaque classe de produit sans tenir compte des autres composants ou de la date de construction.

2- **Spécialisation par ajout d'une contrainte temporelle**. Pour chaque règle hors-contexte générée par l'étape précédente, nous ajoutons le chemin de propriété nécessaire pour atteindre l'année de construction. A partir du produit cible P : $has\_location(S, L), contain(L, P), has\_structure(B, S)$, $has\_year(B, Y)$. Le prédicat *SWRL :lessThanOrEqual(Y, y)* (pour une règle qui conclut sur "positif") ou *SWRL :greaterThanOrEqual(Y, y)* (pour "négatif"), est également ajouté pour comparer l'année de construction Y à une année de référence y qui maximise la confiance et préserve $hc \geq minHc$.

Par exemple, si la règle R1 suivante est générée par la première étape :

R1 : *enduit(P), has_diagnostic_characteristic(P, D)* $\rightarrow$ *has_diagnostic(D, "positif")*

Cette règle peut être spécialisée de la façon suivante :

R2 : *enduit(P), has_location(S, L), contain(L, P), has_structure(B, S), has_year(B, Y), SWRL :lessThanOrEqual(Y, 1980), has_diagnostic_characteristic(P, D)* $\rightarrow$ *has_diagnostic(D, "positif")*

Pour découvrir la meilleure année de référence, CRA-Miner explore les valeurs d'année possibles de la plus récente à la plus ancienne, et considère différemment les règles qui concluent sur "$negatif$" et "$positif$".

La Figure 6.2 montre comment la confiance évolue de 1946 à 1997 pour une règle qui conclut sur "$positif$" et pour une classe de produit. Quand l'année de référence diminue, le *head coverage* hc diminue et la confiance $conf$ augmente. Pour couvrir le nombre maximum de diagnostics en maximisant la confiance, l'exploration s'arrête quand $hc < minHc$ (c.-à-d. 1966 sur la figure 6.2). La dernière année explorée telle que $hc \geq minHc$ et telle que la confiance reste maximum (c.-à-d. 1970 sur la figure 6.2) est choisie. Un processus similaire, mais symétrique est appliqué pour choisir y pour les règles concluant sur "$negatif$".

3- **Spécialisation par localisation et/ou par structure** (prédicat 'Location' et prédicat 'Structure').

Les hiérarchies de localisations et structures sont explorées pour spécialiser les règles générées en étape 1 et 2 avec des composants de bâtiment spécifiques qui contiennent le produit cible P.

Par exemple, la règle R1 peut être spécialisée en spécifiant que la localisation est un mur et que la structure est un balcon.

R3 : *enduit(P), mur(L), balcon(S), has_location(S, L), contain(L, P), has_structure(B, S), has_year(B, Y), SWRL :less-*

ThanOrEqual(Y, 1980), has_diagnostic_characteristic(P, D) $\rightarrow$ *has_diagnostic(D, "positif")*

4- **Enrichissement par la région** Toutes les règles générées peuvent être enrichies par la propriété 'has_region' qui représente la région dans lequel le bâtiment est situé.

5- **Spécialisation en ajoutant d'autres composants**. Dans cette étape, de nouvelles propriétés sont ajoutées qui représentent des produits spécifiques frères, des localisations spécifiques frères ou des structures spécifiques frères : $contain(L, P_i)$ et $C_p(P_i)$ où i varie de 0 à $maxSiblingP$ et C_p est une feuille de la hiérarchie de produits, puis $has\_location(S, L_j), C_l(L_j)$ où j varie de 0 à $maxSiblingL$ et C_L est une feuille de la hiérarchie des localisations), et $has\_structure(S, L_j), C_l(L_j)$ où j varie de 0 à $maxSiblingS$ et C_S et une feuille de la hiérarchie des structures.

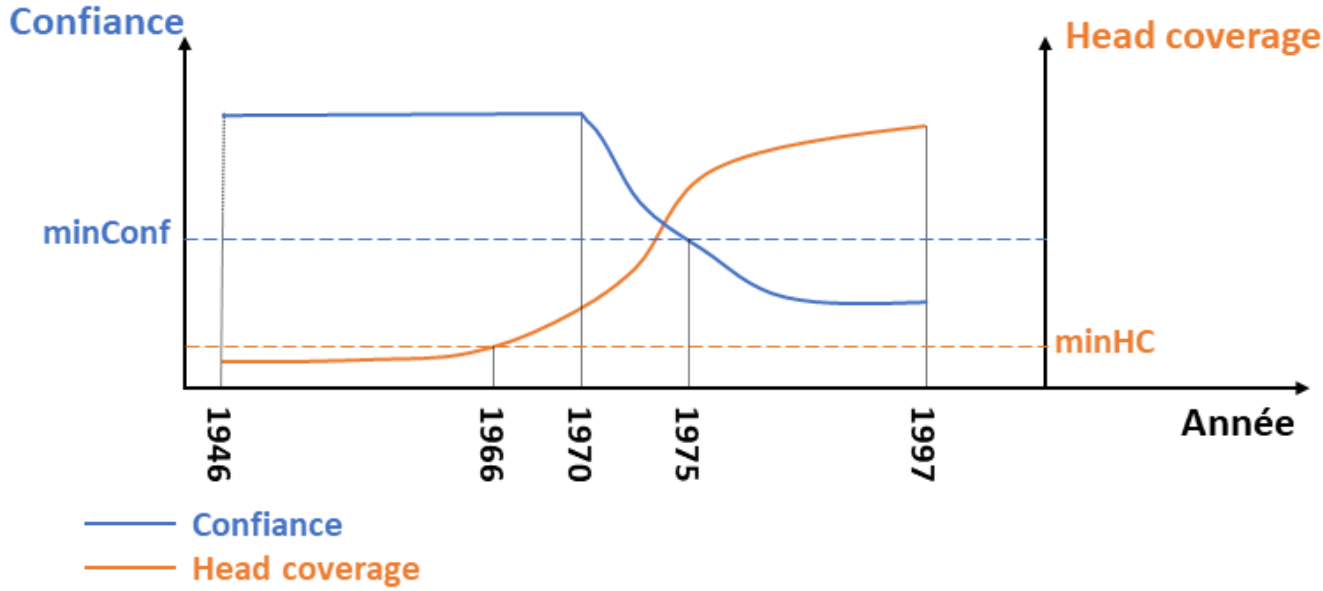


FIGURE 6.2 – Évolution de la confiance et du *head coverage* d'une règle concluant sur "positive"

Chaque spécialisation d'une règle est testée : une spécialisation est conservée si elle satisfait hc et améliore la confiance de la règle dont elle est issue. À la fin de l'algorithme, nous supprimons toutes les règles dont la confiance est inférieure à $minConf$ (la confiance étant non anti-monotone, l'élagage sur la confiance ne peut avoir lieu qu'a posteriori).

6.2.5 Représentation algorithmique de CRA-Miner

CRA-Miner a besoin comme entrée d'un graphe de connaissance K, de la règle la plus générale : $\top = product(P), has\_diagnostic\_characteristic(P, D) \rightarrow has\_diagnostic(D, Valeur)$, un ensemble de seuils pour la qualité des règles (minConf pour la confiance et minHC pour le *head coverage*) et un ensemble de seuils pour limiter l'exploration de l'espace de recherche (maxSiblingsP, maxSiblingsL et maxSiblingsS pour le nombre de produits,

Algorithme 1 : CRA-Miner

Data : GC K, pattern $\top$, minConf, minHC, maxSiblingsP, maxSiblingsL, maxSiblingsS
Result : ruleSet

```
1  ruleSet = {};
2  ruleSet_P = SpecializeProduct(k, ⊤, minHC);
3  ruleSet_PT = TemporalSpecialization(k, ruleSet_P, minConf, minHC);
4  ruleSet_subP = SpecializeProductHierarchy(k, ruleSet_P, minConf, minHC);
5  ruleSet_subPT = TemporalSpecialization(k, ruleSet_subP, minConf, minHC);
6  ruleSet_PT.update(ruleSet_subPT);
7  ruleSet_L = SpecializeLocation(k, ruleSet_PT, minConf, minHC);
8  ruleSet_LT = TemporalSpecialization(k, ruleSet_L, minConf, minHC);
9  ruleSet_subL = SpecializeLocationHierarchy(k, ruleSet_L, minConf, minHC);
10 ruleSet_subLT = TemporalSpecialization(k, ruleSet_subL, minConf, minHC);
11 ruleSet_LT.update(ruleSet_subLT);
12 ruleSet_S = SpecializeStructure(k, ruleSet_LT, minConf, minHC);
13 ruleSet_ST = TemporalSpecialization(k, ruleSet_S, minConf, minHC);
14 ruleSet_Comp = ComponentEnrichment(k, ruleSet_ST, minConf, minHC);
15 ruleSet_Sib = SiblingsEnrichment(k, ruleSet_Comp, minConf, minHC, maxSiblingsP, maxSiblingsL,
     maxSiblingsS);
16 ruleSet_Sib.update(ruleSet_Comp);
17 ruleSet_Final = TemporalSpecialization(k, ruleSet_Sib, minConf, minHC);
18 ruleSet.update(ruleSet_Final);
```

localisations et structures co-localisés). Le résultat attendu de CRA-Miner sera un ensemble de règles “ruleSet” qui conclurent sur la présence ou l’absence de l’amiante dans les produits.

CRA-Miner démarre avec un ruleSet initialement vide (ligne 1). La première étape sera la spécialisation du $\top$ avec une sous-classe de produit (ligne 2) et l’instanciation de $Valeur$ dans la tête de la règle par 0 ou 1. La condition qui doit être satisfaite pour garder les règles générées est $hc \geq minHC$. Cette condition est respectée pour toutes les fonctions suivantes. La ligne 3 montre l’implémentation de l’étape de la spécialisation par ajout d’une contrainte temporelle, c.-à-d. nous ajoutons à chaque règle découverte une contrainte temporelle et nous gardons que celle avec $hc \geq minHC$. Par la suite, la spécialisation temporelle va être appliquée après les prochaines étapes pour toutes les nouvelles règles (lignes 5, 8, 10, 13 et 17). Pour les prochaines étapes, si la confiance d’une règle trouvée $= 1$ nous arrêtera l’exploration pour cette règle. A partir de cette étape, nous gardons les règles les plus spécifiques seulement si elles améliorent la confiance. La ligne 4 montre l’exploration *top-down* de la hiérarchie des produits (par exemple : “revêtement de mur” et une sous-classe de “revêtement” qui apparaisse dans la hiérarchie de “revêtement”) pour trouver des nouvelles règles. Ensuite, une spécialisation temporelle sera appliquée à ces nouvelles règles (ligne 5). Nous avons implémenté l’étape de la spécialisation par localisation dans la ligne 7. Nous ajoutons ensuite une contrainte temporelle aux nouvelles règles dans la ligne 8. L’exploration de la hiérarchie de localisation est implémentée dans la ligne 9 suivie par une spécialisation temporelle pour les nouvelles règles résultantes (ligne 10). Ensuite nous spécialisons les règles avec la structure (ligne 12) et nous appliquons également pour les nouvelles règles une spécialisation temporelle (ligne 13). Nous n’explorons pas la hiérarchie de la structure puisqu’elle est jugée sans influence par l’expert. L’enrichissement par la région est réalisé

par la fonction “ComponentEnrichment” (ligne 14) qui peut aussi enrichir les règles avec d’autres composants que la région comme le type du bâtiment. La dernière étape est l’enrichissement par des composants co-localisés (les produits co-localisés, les localisations co-localisées et les structures co-localisées). La fonction “SiblingsEnrichment” (ligne 15) utilise un ensemble de seuils de composants co-localisés pour limiter le nombre de ces composants qui sont permet d’être ajoutés à la règle, c.-à-d. le nombre de produits co-localisés ajoutés ne doit pas dépasser “maxSiblingsP”, même chose pour le nombre de localisations et de structures co-localisées, il ne doit pas dépasser “maxSiblingsL” et “maxSiblingsS” respectivement. Les nouvelles règles découvertes par les deux dernières fonctions seront enrichies par une contrainte temporelle (ligne 17). Noté bien que si la spécialisation temporelle est appliquée à une règle qui contient déjà une contrainte temporelle, la fonction va recalculer l’année de référence pour assurer qu’elle est la meilleure année qui maximise la confiance. A la fin nous mettrons à jour l’ensemble de résultats “ruleSet” avec les ensembles de règles qui résultent de chaque étape durant la procédure de l’induction des règles par CRA-Miner.

6.3 Complexité de l’espace de recherche

Le nombre de règles générées et testées est principalement impacté par la spécialisation temporelle (dans le pire des cas, tout l’intervalle de temps sera testé) et l’ajout de composants co-localisés (dans le pire des cas, toutes les combinaisons possibles de composants co-localisés seront être vérifié). Cependant, CRA-Miner peut être parallélisé puisque chaque règle peut être spécialisée indépendamment des autres.

Dans cette section, nous allons montrer et discuter la complexité temporelle de chaque étape de CRA-Miner. Notez que CRA-Miner est parallélisé, donc chaque règle est spécialisée indépendamment des autres.

Dans les équations de complexité suivantes, b représente le nombre de processeurs utilisés pour exécuter l’algorithme en parallèle. D’autre part, n représente la taille d’entrée à chaque étape, il représente le nombre de conclusions pour l’étape 1, mais pour les étapes 2, 3, 4 et 5 il représente la taille de l’ensemble de règles actuel à spécialiser.

Remarque : Pour chaque étape, nous montrons la complexité de l’espace de recherche en prenant en compte la parallélisation $T(n)$, mais aussi la taille totale de l’espace de recherche $O(n)$ qui ne considère pas la parallélisation.

1- **Spécialisation de $\top$ en utilisant des sous-classes de produit** : Le but de cette étape est de spécialiser le $\top$ par les différents types de produits dans le GC, sachant que cette spécialisation devrait être pour toutes les classes de règles (c.-à-d. pour toutes les conclusions par exemple dans le cas de l’amiante nous possédons deux classes : amianté et non amianté), le nombre de conclusions va être représenté par n dans la formule suivante :

$$O(n) = pn$$

$$T(n) = \lceil \frac{pn}{b} \rceil$$

Avec :

p :le nombre de types de produits dans le GC.

2- **Spécialisation par ajout d'une contrainte temporelle** : Afin de trouver l'année de référence qui maximise la confiance de la règle, nous testons autant que nécessaire les années de manière séquentielle, dans le pire des cas, nous allons parcourir l'intervalle de temps entier. Ce qui rend cette étape très coûteuse est la longueur de l'intervalle de temps et la complexité de la requête SPARQL de comparaison.

$$O(n) = tn$$

$$T(n) = t\lceil \frac{n}{b} \rceil$$

Avec :

t : la longueur de l'intervalle de temps à tester (en ne considérant que les années).

3- **Spécialisation par localisation et/ou par structure** : Les types de localisation et de structure concernés par cette étape sont les types qui possèdent au moins un individu en relation avec le produit décrit par la règle (c.-à-d. nous testons que les localisations et les structures qui contiennent le produit mentionné dans la règle).

$$O(n) = lsn$$

$$T(n) = \lceil \frac{lsn}{b} \rceil$$

Avec :

l : le nombre de types de localisation.

s : le nombre de types de structure.

4- **Enrichissement par la région** :

$$O(n) = rn$$

$$T(n) = \lceil \frac{rn}{b} \rceil$$

Avec :

r : le nombre de régions.

5- **Spécialisation en ajoutant d'autres composants** : Le pire de cas de cette étape est de tester toutes les combinaisons possibles pour les composants co-localisés.

$$O(n) = C_X^x C_L^l n$$

$$T(n) = C_X^x C_L^l \lceil \frac{n}{b} \rceil$$

Avec :

x : le nombre maximum de produits frères à vérifier (maxSiblingsProduct).

y : le nombre max de localisations frères à vérifier (maxSiblingsLocation).

X : le nombre de types de produits frères.

Y : le nombre de types de localisations frères.

Comme nous pouvons le constater à partir de ces équations, le temps de génération et de test des règles ainsi que leur nombre sont principalement impactés par la spécialisation temporelle et l'ajout de composants co-localisés.

6.4 Expérimentations

Nous avons évalué notre approche sur un GC qui a été alimenté à partir d'un ensemble de documents de diagnostic fournis par le CSTB (le même ensemble utilisé dans les expérimentations du chapitre précédent). Ce GC contient 51970 triplets qui décrivent 2998 instances de produit, 341 localisations, 214 structures et 94 bâtiments. L'année de construction de ces bâtiments varie entre 1948 et 1997. Nous avons 1525 produits qui contiennent de l'amiante et 1473 produits sont sans amiante. Toutes les expérimentations ont été réalisées sur un serveur doté de 80 processeurs physiques (Intel Xeon E7-4830 2,20 GHz) et 528 Go de RAM.

Le but de ces expérimentations est (1) d'apprendre des règles sur un sous-ensemble de diagnostics et d'étudier la qualité de la prédiction qui peut être faite sur les produits restants (2) comparer les résultats de notre approche à une approche naïve qui n'utilise que la classe de produit et l'année de construction (baseline) (3) comparer les résultats de notre approche avec les deux approches d'exploration de règles AMIE3 [27] et TILDE [2] (4) comparer nos résultats avec l'approche hybride présentée dans le chapitre précédent qui calcule la probabilité amiante en utilisant des ressources externes (ANDEVA [1] et INRS [2]).

Pour évaluer notre approche, nous avons divisé les données de GC en 3 tiers, et nous avons effectué une validation croisée, c.-à-d. à chaque test, nous utilisons 2 tiers pour apprendre des règles et le tiers restant pour tester les règles apprises. Comme nous avons de nombreuses classes de produits différentes, nous fixons un seuil bas de *head coverage* à $minHC = 0,001$ pour observer le plus de règles possible. Ensuite, nous avons évalué

1. Association nationale pour la défense des victimes de l'amiante http://andeva.free.fr/expositions/gt_expos_produits.htm
2. Institut National de Recherche et de Sécurité http://www.inrs.fr/media.html?refINRS=ED%201475

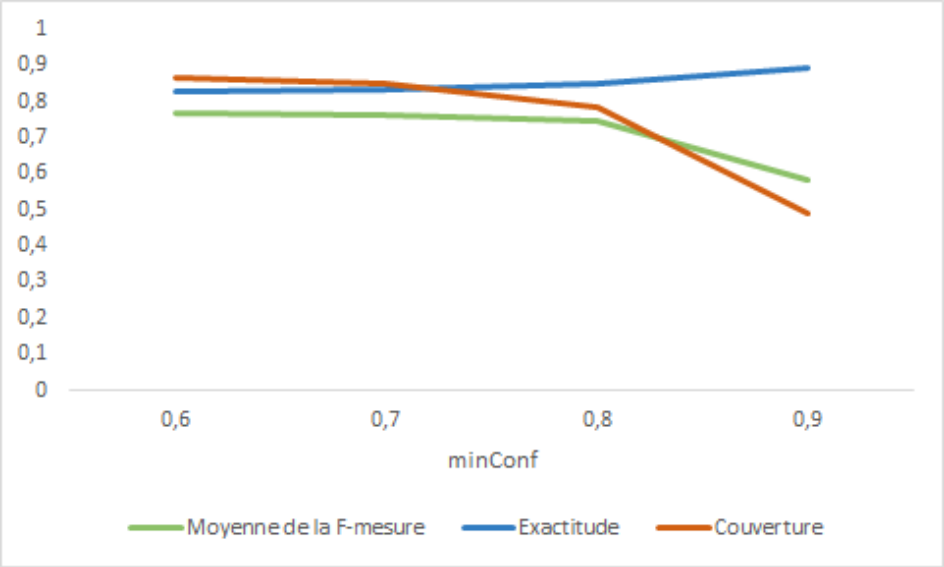


FIGURE 6.3 – Résultats de CRA-Miner selon le seuil $minConf$

les résultats lorsque $minConf$ varie de $0,6$ à $0,9$ en utilisant les mesures classiques de précision, de rappel, de F-Measure et d'*accuracy*. Le nombre maximum de frères a été fixé à 0 pour les structures et à 3 pour les localisations et les produits par l'expert.

Le tableau 6.1 montre que CRA-miner découvre en moyenne 75 règles. Les résultats montrent que les composants co-localisés sont efficacement exploités pour prédire la présence d'amiante : 29 règles en moyenne impliquent au moins un produit frère (maximum 2 produits frères) et 17 règles impliquent des localisations frères (maximum 3 localisations). De plus, les résultats montrent que CRA-miner a découvert 14 règles en moyenne qui exploitent une contrainte temporelle.

Nous avons adhéré à une approche pessimiste qui choisit de classer un produit comme positif si au moins une règle conclut qu'il contient de l'amiante.

La Figure 6.3 présente la moyenne de la F-mesure, l'*accuracy* et la couverture (c'est-à-dire le rapport des produits pouvant être classés dans l'ensemble de test) lorsque $minConf$ varie. Ces résultats suggèrent que lorsque le seuil $minConf$ augmente, la précision augmente, mais la couverture des données diminue. La meilleure moyenne F-mesure $0,77$ (moyenne entre la F-mesure positive et négative) est obtenue pour un $minConf$ fixé à $0,6$. Avec un tel seuil, nous pouvons décider pour 87% des échantillons de test. La Figure 6.4 détaille les résultats (TN, TP, FN, FP, négatif non classé : UN, positif non classé : UP) ainsi que le nombre de produits qui ont été classés comme positifs et négatifs par différentes règles (doubles décisions) . Plus précisément, les vrais positifs TP (respectivement vrais négatifs TN) sont les produits contenant de l'amiante classés par les règles découvertes comme positifs (respectivement négatifs). Les faux positifs FP (respectivement faux négatifs FN) sont les produits sans amiante classés par les règles comme positifs, tandis que les produits non classés sont soit positifs (UP) soit sans amiante (UN) dans le GC. Cette figure montre qu'un seuil fixé à $0,6$ conduit à une moyenne de seulement 82 de décisions

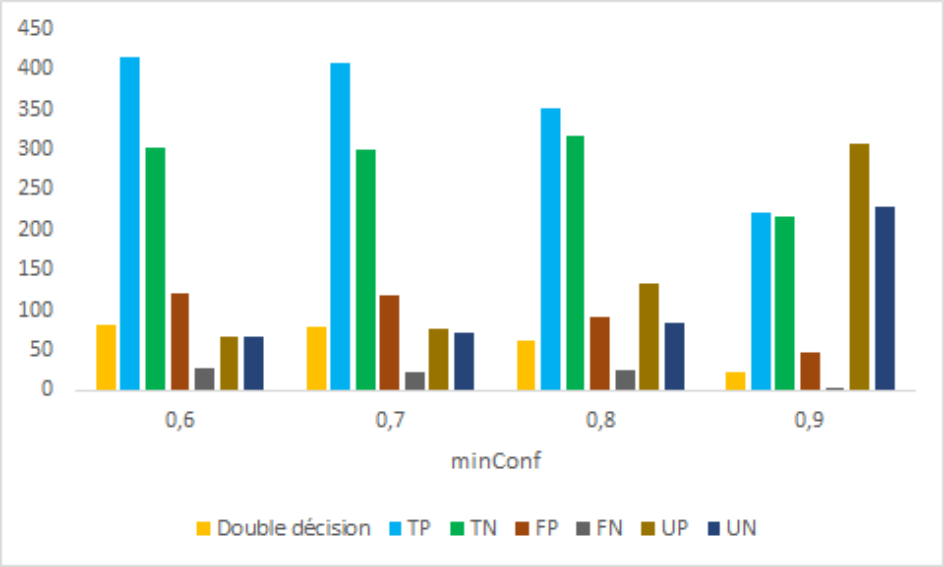


FIGURE 6.4 – Résultats détaillés de CRA-Miner selon les seuils $minConf$

contradictoires pour des échantillons de test qui décrivent un millier de produits.

Nous avons comparé l'approche contextuelle CRA-Miner avec une baseline non contextuelle qui est uniquement basée sur la classe de produits à exploiter. Cette baseline nous permet d'estimer les avantages de prendre en compte la hiérarchie des produits et le contexte dans lequel ils ont été utilisés (c.-à-d. le type de la localisation et les autres produits utilisés dans la même localisation). La figure 6.5 montre que la F-mesure et la couverture sont plus faibles pour l'approche de base (la baseline) quel que soit le seuil $minConf$ qui varie de $0,6$ à $0,9$. En particulier, le tableau 6.1 montre que la baseline ne classe que 46% des échantillons de test et obtient une moyenne de F-mesure de $0,55$.

En effet, CRA-miner permet de découvrir des règles complexes telles que :

enduitsabasedeplatreoudecimentprojeteslissesoutaloches(?P), has_location(?S, ?L), contain(?L, ?P), enduitsderagreagedebullagelissage(?P2), contain(?L, ?P2), has_structure(?B, ?S), has_year(?B, ?Y), has_diagnostic_characteristic(?P, ?D) , lessThanOrEqual(?Y, "1991-01-01T00 :00 :00") $\rightarrow$ *has_Diagnosis(?D, "positive")*

Nous avons comparé nos résultats obtenus avec AMIE3 [27] en utilisant les mêmes seuils de $minConf$ et $minHC$, et en fixant le nombre de prédicats des règles recherchées à $l = 4$ et $l = 6$ (cf .table 6.1)[3]. Notre approche permet d'obtenir une meilleure F-mesure que ce qui est obtenu avec [27] (0,77 contre 0,73 pour $l = 6$, le l spécifié étant le nombre de prédicats permettant à AMIE3 d'obtenir les meilleurs résultats en termes de F-mesure et précision). AMIE3 a pu découvrir 91 règles (75 avec notre approche) ce qui lui permet de couvrir 100% des données de test (87% avec notre approche). En revanche, il obtient une précision moindre (0,74 contre 0,83 avec CRA-Miner). Cette couverture importante s'accompagne de nombreuses doubles décisions (277). L'approche est pessimiste (c.-à-d. si un produit est associé à deux décisions différentes, elle considère le produit comme contenant

3. Malgré le fait que AMIE3 est utilisé pour rechercher uniquement des règles concluantes sur $has\_Diagnosis$, une longueur > 6 ne donne pas de résultats en moins de trois semaines.

de l'amiante), AMIE3 trouve plus de TP (473 contre 415 avec CRA-Miner), mais presque deux fois plus de FP (226 contre 121 avec CRA-Miner), et les TN sont également moins nombreux (seulement 264 contre 303 avec CRA-Miner). Avoir un contexte sémantique et pouvoir représenter des intervalles de temps permet de découvrir des règles qui impliquent plus d'atomes tout en améliorant leur lisibilité pour un expert du domaine (plus précisément, une règle peut être définie pour un intervalle de temps alors que AMIE3 ne peut générer que des règles portant sur une année spécifique).

De plus, nous avons testé notre biais de langage à l'aide du système TILDE [2] qui génère des arbres de décision relationnels qui permettent de représenter des biais de langage complexes émulant des langues cibles similaires (contexte relationnel et valeurs maxSibling) et de gérer (bien que pas de manière optimale) une hiérarchie de types. Le contexte relationnel utilisé est légèrement différent, n'imposant qu'au moins un type instancié dans le contexte (pas nécessairement le produit). Cette stratégie *top-down* obtient par définition une couverture de 100% de produits sans double décision, mais conduit à une précision moindre pour les exemples positifs et négatifs. En effet, il obtient plus de FP et FN puisque la dernière règle générale classe tous les individus restants non classés comme positifs ou négatifs, quelle que soit ses descriptions. CRA-Miner n'a pas pu classer tous les exemples (87%), mais la précision obtenue est plus élevée (0,83 contre 0,79 pour le TILDE). Compte tenu de la stratégie de TILDE, il n'a pas été possible d'utiliser les inégalités sur les années, car introduire cette possibilité donne la possibilité d'apprendre des intervalles fermés sur les années et un surapprentissage difficilement contrôlable.

Nous avons également comparé CRA-Miner à l'approche hybride présentée dans le chapitre précédent. Cette approche utilise deux ressources externes qui décrivent des produits commercialisés contenant de l'amiante pendant au moins une période pour calculer une probabilité basée sur la classe de produits et l'année de construction. Le tableau 6.1 montre que l'approche hybride obtient une F-mesure et une précision plus élevées en particulier pour les produits positifs ($0,94$ contre $0,79$ pour CRA-Miner). Cela peut s'expliquer par les informations supplémentaires fournies par les ressources Web qui se concentrent sur les produits positifs. Cependant, CRA-Miner pourrait couvrir davantage d'échantillons de données (87% contre 83% pour l'hybride). En effet, l'approche hybride ne pouvait trancher sur un produit si sa classe n'était pas mentionnée dans les ressources externes.

Ces expérimentations ont d'abord montré que tous les prédicats du contexte retenus par l'expert sont pertinents pour classer les produits. En effet, la baseline obtient un rappel très faible, et les résultats montrent que tous les prédicats ont été utilisés dans au moins une règle. La comparaison avec les deux autres systèmes d'extraction de règles disponibles montre que CRA-Miner obtient les meilleures valeurs de précision, avec une valeur de couverture plus faible, mais toujours élevée (87%). Comme attendu, les expérimentations montrent également que l'utilisation de ressources externes sur des produits commercialisés contenant de l'amiante peut conduire à des décisions plus précises. Cependant, ce type de ressource est incomplet, et la couverture obtenue est plus faible. Puisqu'il est plus important de détecter les exemples positifs que négatifs, nous avons choisi d'appliquer une stratégie pessimiste, et les résultats montrent que nous obtenons un meilleur rappel pour les exemples positifs que pour les exemples

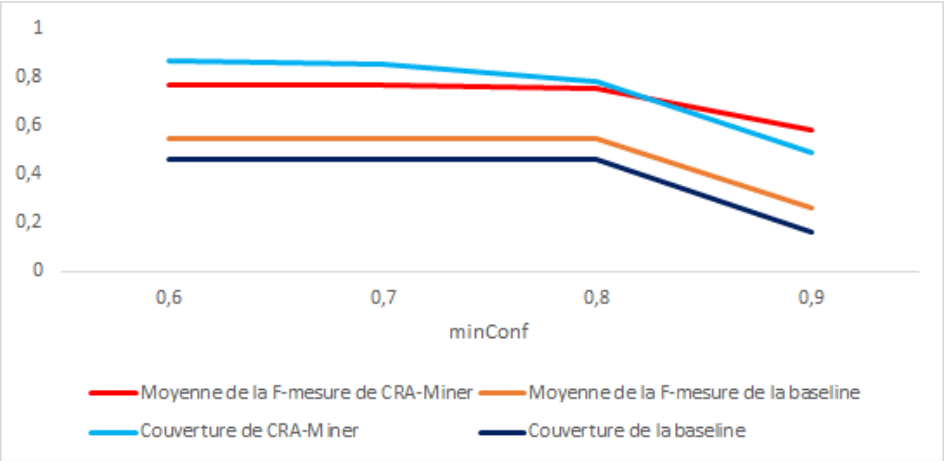


FIGURE 6.5 – Comparaison entre les approches contextuelles et non contextuelles selon les seuils de minConf

Classification du système	Systèmes d'extraction de règles					Systèmes basés sur des ressources externes
Système	CRA-Miner	AMIE3 $l = 4$	AMIE3 $l = 6$	TILDE	Baseline	Approche hybride
# règles	75	45	91	34	24	/
Double décision	82	50	277	0	0	0
TP	415	381	473	431	146	465
TN	303	288	264	358	257	348
FP	121	146	226	128	30	16
FN	28	74	32	87	24	5
UP	66	54	3	0	338	38
UN	67	58	0	0	204	127
Pos. précision	**77%**	72%	68%	**77%**	83%	97%
Pos. rappel	82%	75%	**93%**	83%	29%	92%
Pos. F-mesure	0,79	0,73	0,79	**0,80**	0,43	0,94
Neg. précision	**92%**	80%	89%	80%	91%	99%
Neg. rappel	62%	59%	54%	**74%**	52%	71%
Neg. F-mesure	0,74	0,68	0,67	**0,77**	0,66	0,83
Moy. F-mesure	0,77	0,71	0,73	**0,79**	0,55	0,89
Accuracy	**0,83**	0,75	0,74	0,79	0,88	0,97
Couverture	87%	89%	**100%**	**100%**	46%	83%

TABLE 6.1 – Comparaison entre CRA-Miner, AMIE3 avec $l = 4$ et $l = 6$ ($minHC = 0,001$, $minConf = 0,6$), TILDE, baseline, et l'approche hybride

négatifs. Cependant, ce choix affecte la précision des points positifs et d'autres stratégies pourraient être envisagées (par exemple, des stratégies de vote, des règles ordonnées en fonction de leur sémantique et/ou de leur confiance). Une autre possibilité consiste à utiliser un seuil de confiance plus élevé pour les négatifs. Les résultats ont montré que lorsque la valeur de confiance est fixée à $1,43\%$ des négatifs peuvent encore être découverts avec un seul faux négatif parmi 210 décisions ($99,52\%$ de précision).

6.5 Combinaison de CRA-Miner avec l'approche hybride

Nous avons montré dans le chapitre précédent que l'approche hybride basée sur l'exploitation de ressources externes obtenait une très bonne précision, mais que la couverture pouvait être améliorée. L'approche CRA-Miner, elle, permet de couvrir plus d'exemples, mais sa précision est plus faible.

Dans ce chapitre nous proposons et discutons de différentes stratégies permettant de combiner ces deux approches dans le but d'améliorer la couverture des données tout en maintenant une bonne précision. Pour cela, nous avons analysé les ensembles d'instances de produit classés comme négatifs, positifs ou non classés trouvés par chaque approche et comparé les résultats en nous basant sur différentes méthodes.

6.5.1 Comparaison et discussion

Dans cette section, nous allons comparer et analyser les résultats obtenus par CRA-Miner et par l'approche hybride en considérant l'ensemble des produits utilisés précédemment dans les jeux de données test (999 produits en moyenne). Nous récapitulons dans la Table 6.2 le nombre moyen de produits classés comme étant négatifs, positifs ou non classés (c.-à-d. notés /) trouvés par CRA-Miner ou par l'approche hybride. Pour chaque ligne, l'information de présence d'amiante dans les diagnostics est présente dans la colonne Diagnostic. Par exemple, la première ligne montre qu'il y a 410 produits positifs qui sont bien classés à la fois par CRA-Miner et par l'approche hybride.

La Table 6.2 montre que lorsque CRA-Miner et l'approche hybride aboutissent à la même décision pour un produit, cette décision est correcte à l'exception des 2 décisions apparaissant en Ligne 10. Ces résultats montrent également que l'approche hybride et CRA-Miner peuvent aboutir à des décisions contradictoires. Sur les 999 produits, 134 décisions sont contradictoires. L'analyse plus précise de ces cas présentés dans les lignes 2 et 3 pour les positifs, et dans les lignes 11 et 12 pour les négatifs, montre que l'approche hybride obtient la bonne classe pour tous les produits sauf pour un seul produit (ligne 2). En cas de décision contradictoire, l'analyse montre qu'il faut donc privilégier l'approche hybride. Il existe 213 produits pour lesquels seule l'une des deux approches peut décider. Plus précisément, CRA-Miner peut classifier 123 produits pour lesquels aucune décision n'est prise par l'approche hybride, dont 106 sont classifiés correctement (86 % d'*accuracy*). En effet, les produits qui n'ont jamais été amiantés ne sont pas décrits dans les ressources externes. Aussi, la première approche ne permet pas de

	Diagnostic	CRA-Miner	Approche hybride	Moyenne
1	Positif	Positif	Positif	410
2	Positif	Positif	Négatif	1
3	Positif	Négatif	Positif	17
4	Positif	Négatif	Négatif	0
5	Positif	Positif	/	4
6	Positif	Négatif	/	10
7	Positif	/	Positif	38
8	Positif	/	Négatif	3
9	Positif	/	/	24
10	Négatif	Positif	Positif	2
11	Négatif	Positif	Négatif	116
12	Négatif	Négatif	Positif	0
13	Négatif	Négatif	Négatif	196
14	Négatif	Positif	/	3
15	Négatif	Négatif	/	106
16	Négatif	/	Positif	14
17	Négatif	/	Négatif	35
18	Négatif	/	/	18

TABLE 6.2 – Combinaison entre CRA-Miner et l'approche hybride

les classifier. Nous pouvons quand même noter que l'approche hybride peut classifier 90 produits non classifiés par CRA-Miner dont 73 sont bien classés (81 % d'*accuracy*). Dans le cas où une seule approche peut décider, l'*accuracy* montre que nous pouvons donc considérer directement cette classification.

6.5.2 Stratégies de combinaison des deux approches

Compte tenu de l'analyse des ensembles des produits bien classés, mal classés ou non classés effectuée, nous proposons une stratégie, notée CRA-Miner+H, permettant de privilégier les décisions prises par l'approche hybride. Les seules décisions de CRA-Miner qui sont considérées sont celles qui concernent des produits pour lesquels l'approche hybride n'a pas pu décider.

Nous avons comparé cette stratégie avec deux autres stratégies permettant de combiner les deux approches :

— La première stratégie est une approche pessimiste qui privilégie les décisions positives : il suffit que l'une des deux approches classifie le produit comme amianté pour décider que celui-ci est amianté.

— La deuxième stratégie privilégie les décisions de CRA-Miner. Les seules décisions de l'approche hybride qui sont considérées sont celles qui concernent des produits pour lesquels CRA-Miner n'a pas pu décider.

Nous montrons dans la Table 6.3 les nouveaux résultats obtenus suite à la combinaison de CRA-Miner et de l'approche hybride. L'utilisation des décisions d'une approche lorsque l'autre n'a pas pu classifier un individu a grandement amélioré la couverture des données (96%). L'application de la stratégie privilégiant les décisions de l'approche hybride permet d'obtenir des précisions très hautes pour les positifs et les négatifs (respectivement 96% et 97%) et aussi une très bonne *accuracy* (0,96). La stratégie pessimiste, même si elle permet d'avoir un rappel

Approche	CRA-Miner	Approche hybride	Combinaison : approche hybride privilégiée CRA-Miner+H	Combinaison : Approche pessimiste	Combinaison : CRA-Miner privilégié
TP	415	465	469	470	453
TN	303	348	454	338	338
FP	121	16	19	135	31
FN	28	5	15	14	135
UP	66	38	24	24	24
UN	67	127	18	18	18
Pos. précision	77%	**97%**	96%	78%	94%
Pos. rappel	82%	92%	92%	**93%**	74%
Pos. F-mesure	0,79	**0,94**	**0,94**	0,85	0,83
Neg. précision	92%	**99%**	97%	96%	71%
Neg. rappel	62%	71%	**92%**	69%	87%
Neg. F-mesure	0,74	0,83	**0,94**	0,80	0,78
Moy. F-mesure	0,77	0,89	**0,94**	0,83	0,81
Accuracy	0,83	**0,97**	0,96	0,84	0,83
Couverture	87%	83%	**96%**	**96%**	**96%**

TABLE 6.3 – Résultats de la combinaison entre CRA-Miner et l'approche hybride, en considérant 3 stratégies

légèrement plus élevé pour les produits amiantés (93% contre 92%), obtient une précision beaucoup plus faible pour ces produits (seulement 78%). La stratégie privilégiant CRA-Miner obtient l'*accuracy* la plus faible et ce résultat est dû à une précision beaucoup plus faible pour les produits non amiantés (71%).

La Table 6.4 compare les résultats de CRA-Miner+H avec les approches TILDE et AMIE3. Les résultats montrent que CRA-Miner+H obtient une bien meilleure *accuracy* (0,96 contre 0,75 pour AMIE3 et 0,79 pour TILDE) pour une couverture qui est maintenant proche de 100% qui est atteinte par les deux autres approches. De plus, le rappel des produits amiantés est presque aussi élevé que celui d'AMIE3 avec une précision beaucoup plus forte pour les positifs (96%) et les négatifs (97%).

6.6 Conclusion

Dans ce chapitre, nous avons présenté l'approche de découverte de règles CRA-Miner qui permet de prédire la présence d'amiante dans les produits en fonction d'un contexte sémantique, d'heuristiques dédiées aux relations partie-tout et de contraintes calculées sur des valeurs numériques qui représentent des informations temporelles. Les expérimentations montrent que nous pouvons obtenir une meilleure précision et *accuracy* que deux autres systèmes d'extraction de règles et une meilleure couverture que l'approche hybride basée sur des ressources externes.

La combinaison des résultats de CRA-Miner et de l'approche hybride en utilisant une stratégie privilégiant les décisions prises par l'approche hybride permet de proposer des prédictions pour beaucoup plus d'individus. De plus, les résultats de la combinaison sont plus précis que ceux obtenus par CRA-Miner, que ce soit pour les produits

Système	AMIE3 $l = 4$	AMIE3 $l = 6$	TILDE	CRA-Miner+H
TP	381	473	431	469
TN	288	264	358	454
FP	146	226	128	19
FN	74	32	87	15
UP	54	3	0	24
UN	58	0	0	18
Pos. précision	72%	68%	77%	**96%**
Pos. rappel	75%	**93%**	83%	92%
Pos. F-mesure	0,73	0,79	0,80	**0,94**
Neg. précision	80%	89%	80%	**97%**
Neg. rappel	59%	54%	74%	**92%**
Neg. F-mesure	0,68	0,67	0,77	**0,94**
Moy. F-mesure	0,71	0,73	0,79	**0,94**
Accuracy	0,75	0,74	0,79	**0,96**
Couverture	89%	**100%**	**100%**	96%

TABLE 6.4 – Comparaison entre AMIE3 avec $l = 4$ et $l = 6$, TILDE et CRA-Miner+H

détectés comme amiantés ou non-amiantés et cela découle de la préférence donnée à la décision obtenue par l'approche hybride dans le cas des décisions contradictoires, cette précision étant atteinte grâce à son utilisation des ressources externes.

Chapitre 7

Conclusion générale

Nous présentons, dans ce chapitre, un résumé des principales contributions de cette thèse puis quelques perspectives à court terme et à long terme.

7.1 Contributions de la thèse

7.1.1 L'ontologie ASBESTOS

Nous avons créé la partie haute de l'ontologie ASBESTOS qui contient les concepts et les propriétés nécessaires pour représenter les bâtiments, leurs composants (c.-à-d. structures, localisations et produits) et les fichiers sources à l'origine de ces informations. L'ontologie permet également de conserver les résultats des diagnostics amiante réalisés sur les produits d'un bâtiment quand ils existent. Nous avons utilisé les ressources documentaires du CSTB pour enrichir cette ontologie avec de nouveaux sous-concepts des classes Structures, localisations et produits et peupler l'ontologie avec les instances apparaissant dans les documents.

L'extraction des données suit deux méthodes qui diffèrent selon le type de la ressource (diagnostic ou projet homologué). La première méthode d'extraction est adaptée aux diagnostics qui se présentent sous forme de données tabulaires dans lesquelles les termes sont bien structurés dans des cellules de tableau. La deuxième méthode d'extraction exploite les caractéristiques régulières des données semi-structurées des projets homologués contenant les descriptions des termes et les relations entre les termes pour les extraire.

Cette ontologie enrichie et peuplée utilise le vocabulaire défini par la norme NF X46-020 et sa partie haute a été validée par un expert du CSTB. Elle comporte 71 classes, 12 propriétés et 94 instances de bâtiments décrits par 51970 triplets.

Deux modules peuvent venir compléter l'ontologie selon l'approche de prédiction d'amiante utilisée. Un premier module permet de conserver des informations issues de ressources externes et décrivant des ensembles de pro-

duits commercialisés avec les différentes périodes dans lesquelles ils ont été amiantés ainsi que les résultats d'une approche de prédiction utilisant ces données. Un deuxième module permet de conserver des résultats de prédiction réalisés par la deuxième approche.

7.1.2 Deux approches de prédiction de présence d'amiante

L'objectif des deux approches proposées est de répondre au problème de prédiction de présence d'amiante dans les bâtiments. Il existe de nombreux bâtiments qui sont encore amiantés malgré le fait que l'utilisation de ce produit soit interdite depuis 1997. Ces deux approches pourront être utilisées pour aider les experts amiante du CSTB ou d'autres utilisateurs à prioriser les bâtiments et les parties de bâtiments dans lesquels il faut réaliser des diagnostics.

La difficulté majeure est que nous ne disposons que du type de produit (enduit, peinture, etc.) dans la description d'un bâtiment et que nous ne disposons pas de la référence exacte du produit commercialisé utilisé. Ces données peuvent être considérées comme incomplètes dans ce contexte.

Pour pallier cette difficulté, nous avons fait l'hypothèse qu'il était possible d'exploiter les données temporelles concernant la présence d'amiante sur les produits commercialisés pour prédire la présence d'amiante dans une classe de produit utilisée dans un bâtiment construit à une année donnée. Cela a mené au développement d'une première approche de prédiction. La deuxième hypothèse est que le contexte dans le lequel est utilisé le produit peut influencer le choix du produit commercialisé et donc la présence d'amiante dans ce produit. Cette hypothèse nous a conduits à définir une deuxième approche de prédiction.

- **Une approche hybride basée sur des ressources externes**

La première approche que nous avons proposée, appelée approche hybride, utilise les ressources externes fournies par l'INRS et l'ANDEVA pour calculer une probabilité de présence d'amiante en basant sur les données sur les produits commercialisés décrits par ces ressources.

La première difficulté rencontrée avec les ressources externes est que celles-ci donnent parfois des informations conflictuelles et partielles. Elles sont conflictuelles, car une ressource peut indiquer que le produit est amianté tandis que l'autre indique une absence d'amiante, et partielles, car les ressources peuvent mentionner que l'information est non renseignée sur certaines périodes. Pour résoudre ces conflits et prendre en compte l'incertitude résultant de l'information partielle, nous avons choisi de représenter la présence d'amiante pour une référence de produit commercialisé par une probabilité et nous avons adopté une méthode pessimiste pour fusionner les probabilités extraites depuis l'INRS et l'ANDEVA. Les probabilités fusionnées représentent la valeur maximale des probabilités extraites depuis les deux ressources pour la même référence de produit et pour la même période de temps. Les probabilités de présence d'amiante calculées par l'approche hybride basent sur les probabilités fusionnées pour

représenter le nombre de produits commercialisés amiantés pour une période de construction donnée.

La deuxième difficulté réside dans le fait que les ressources externes sont incomplètes, c.-à-d. qu'elles ne décrivent que les produits qui ont été amiantés durant au moins une période durant leurs commercialisations. Si un produit n'a jamais été amianté, il ne sera pas du tout mentionné. Aussi, nous avons évalué la proportion de produits amiantés en utilisant les diagnostics d'amiante du CSTB pour réajuster la probabilité fusionnée grâce à cette proportion ce qui permet de prendre en compte les produits non mentionnés.

Cette approche classifie les produits en deux catégories, ceux avec une forte probabilité d'être amiantés et ceux avec une faible probabilité. La classification se base sur l'apprentissage d'un seuil de classification à partir d'un sous-ensemble de diagnostics du CSTB, ensuite nous comparons les probabilités avec ce seuil pour classifier les produits. Afin de permettre la représentation des nouvelles données des ressources externes, nous avons enrichi l'ontologie avec le module permettant de représenter les données externes d'une part, et les résultats de prédiction d'autre part.

Nous avons évalué l'approche hybride en exploitant le graphe de connaissances enrichi par les ressources externes et par un ensemble de diagnostics du CSTB. Les résultats ont montré que l'approche obtient une très bonne *accuracy*, et un bon rappel en particulier pour les produits amiantés. Les expérimentations ont également montré que si la probabilité calculée devait être réajustée, ce réajustement pouvait être réalisé en utilisant un petit ensemble de diagnostics.

Nous avons également comparé cette approche avec différentes approches naïves, mais également avec deux autres approches d'apprentissage de règles. La première est une approche récente de fouille de règles (AMIE3) et la deuxième est une approche qui construit des arbres de décision sur des données relationnelles pour induire des règles de prédiction (TILDE). Ces expérimentations montrent que l'approche hybride obtient des décisions plus précises. En revanche, compte tenu de l'incomplétude des ressources externes, l'approche hybride ne permet pas de prendre une décision pour les produits appartenant à un type de produit non mentionné dans les ressources externes.

• L'approche de prédiction CRA-Miner

La deuxième approche que nous avons proposée, CRA-Miner, est une approche de découverte de règles de classification qui exploite uniquement les diagnostics existants dans le graphe de connaissance. Elle utilise l'ontologie ASBESTOS et s'inspire des approches de type "Générer et Tester" définies en Programmation Logique Inductive (PLI) pour découvrir des règles représentées en logique du premier ordre qui seront utilisées par le raisonneur pour classifier les produits comme amiantés et non-amiantés.

CRA-Miner permet d'utiliser les heuristiques dédiées aux relations partie-tout qui représentent le contexte dans lequel un produit a été utilisé. De plus, CRA-Miner utilise la sémantique de l'ontologie pour l'exploration *Top-down* de la hiérarchie des concepts et des sous-concepts. CRA-Miner permet également d'utiliser les données temporelles

en générant des prédicats de comparaison de l'année de construction avec des valeurs numériques calculées par CRA-Miner.

Nous avons évalué CRA-Miner en exploitant le graphe de connaissances enrichi par un ensemble de diagnostics du CSTB. Nous avons également comparé CRA-Miner avec l'approche hybride, avec une approche naïve, et avec deux autres approches d'apprentissage de règles. La première est l'approche de fouille de règles (AMIE3) et la deuxième est l'approche TILDE. Les résultats ont montré que CRA-Miner obtient une très bonne *accuracy* et une meilleure précision pour les négatifs ainsi que pour les positifs que les approches n'utilisant pas de ressources externes. En revanche, si CRA-Miner permet de prendre plus de décisions que l'approche hybride, la précision est moins bonne.

• **L'approche combinée CRA-Miner+H**

Nous avons proposé une approche appelée CRA-Miner+H permettant de combiner CRA-Miner et l'approche hybride. Cette combinaison utilise une stratégie qui privilégie les décisions prises par l'approche hybride et qui complète ces décisions par des décisions prises par CRA-Miner.

Les expérimentations de CRA-Miner+H montrent que la couverture de données est renforcée grâce la combinaison des deux approches et que l'*accuracy* reste élevée. Cette combinaison permet d'obtenir des résultats plus précis que l'ensemble des approches et des baselines avec lesquelles nous avons comparé notre approche en classifiant presque autant d'individus que AMIE3 et TILDE.

7.2 Perspectives

Les perspectives que nous allons décrire se divisent en deux catégories, des perspectives à court terme et des perspectives à long terme.

7.2.1 Perspectives à court terme

Les premières perspectives concernent l'interface AsbestosReveal, la représentation des résultats et les interactions possibles avec l'expert. Nous représentons par la suite les perspectives liées à chacune des approches.

• Les deux approches proposées sont implémentées afin d'aider l'expert de CSTB à prioriser les bâtiments dans lesquels il faut faire des diagnostics sur la présence d'amiante. Pour réaliser ceci, nous avons mis en place l'outil "AsbestosReveal" (voir Annexe A) qui inclut les deux approches et une interface simple qui permet à l'expert de lancer les approches et d'afficher les résultats. Cependant les résultats sont soit affichés à l'expert via l'interface sous un format tabulaire, soit enregistrés dans un fichier .owl. L'une des perspectives à court terme de notre travail serait donc de proposer d'autres formats de représentation des données et des résultats pour faciliter leur exploitation (p.

ex. exporter les résultats dans des fichiers sous format XML ou CSV, trier les résultats par certains critères comme l'année, la probabilité, la classe, etc.).

• Une autre perspective que nous allons mettre en oeuvre est de pouvoir expliquer une décision à l'expert via AsbestosReveal. En effet, l'outil devrait être capable de justifier un résultat, mais cette justification varie en fonction de l'approche utilisée. Dans le cas de l'approche hybride, l'interface pourrait afficher les données externes utilisées, ainsi que le calcul détaillé. Dans le cas de CRA-Miner, l'explication peut être constituée de l'ensemble des règles applicables avec leur décision. Dans le cas de la combinaison, l'explication pourra varier en fonction de l'approche privilégiée pour le produit en question, mais l'expert devrait également pouvoir effectuer une comparaison entre les différents résultats de chaque approche pour un produit. Cette perspective nécessite d'étudier en collaboration avec l'expert les éléments d'explications qui lui seraient le plus utiles pour comprendre et éventuellement valider ou invalider un résultat précis ou une règle de décision.

• Nous avons montré dans le chapitre 5 que l'approche hybride classifie les produits soit par le calcul de la probabilité de présence d'amiante ou bien par la propagation des probabilités déjà calculées pour des produits du graphe de connaissance vers de nouveaux produits qui sont de la même classe. Ces règles de propagation exploitent les caractéristiques de l'évolution de la présence d'amiante (p. ex. si une classe de produit n'est plus amiantée en 1982, la règle propage cette information pour un produit dont l'année est postérieure). Comme le calcul est un processus plus coûteux que la propagation des probabilités comme cela est montré dans la section 5.7.5, l'objectif serait de définir la meilleure stratégie pour utiliser les deux méthodes afin d'optimiser et minimiser le nombre de calculs de probabilité effectués. L'idée serait d'itérer une étape d'application des règles de propagation avec une étape de calcul des probabilités pour les produits qui restent sans décision. La difficulté est de choisir les produits pour lesquels le calcul va être réalisé à chaque itération.

• L'approche hybride se base sur l'année, la classe de produit et le coefficient de réajustement pour calculer la probabilité d'un produit appartenant à cette classe d'être amianté. De plus, nous avons montré que cette probabilité est décroissante, c.-à-d. si un produit à une année donnée devient non amianté, il restera non amianté pour toutes les années postérieures. Pour ne pas calculer la probabilité pour chaque produit, nous envisageons de représenter pour chaque classe les intervalles de temps dans lequel un produit peut être considéré comme amianté ou non. Pour cela, il s'agirait de trouver pour chaque type de produit l'année de changement de classification (où le produit passe de l'état amianté à l'état non amianté). Représenter cette année pour tous les types de produits permettrait de réduire considérablement les calculs. Cependant, la recherche de l'année de changement dépend de l'ensemble des diagnostics utilisés dans le réajustement et de ceux utilisés pour trouver le seuil de décision. Si l'on souhaite prendre en compte un nouvel ensemble de diagnostics, l'année devra être mise à jour.

• Nous avons montré dans le chapitre 6 que CRA-Miner découvre des règles SWRL qui sont utilisées par le raisonneur Pellet. Comme Pellet exécute toutes les règles pour saturer le graphe, nous faisons face à deux problèmes. Le premier problème est que nous ne pouvons pas savoir quel est l'ensemble de règles applicables sur

un produit : pour trouver les règles qui peuvent être appliquées, il faudrait réaliser une requête spécifique à chaque prémisse de règles pour découvrir tous les produits concernés. Le deuxième problème est qu'il n'est pas possible d'ordonner les règles en fonction de leur qualité (c.-à-d. leur confiance, leur complexité en termes de nombre de prédicats, niveau de la hiérarchie, etc.). Or, suite à l'application de règles concluant sur la classe amiantée et non amiantée pour le même produit, nous obtenons des produits associés à des décisions contradictoires. Pour résoudre les deux problèmes, il faudrait, dans un premier temps, ordonner les règles selon les critères choisis. Dans un deuxième temps, il faut créer une nouvelle méthode qui est capable d'appliquer séquentiellement les règles selon leurs scores, et éliminer ainsi les doubles décisions (c.-à-d. la nouvelle méthode doit vérifier si un produit est déjà classifié avant d'appliquer une nouvelle règle). Une deuxième possibilité consisterait à adopter d'autres stratégies pour résoudre les doubles décisions que la stratégie pessimiste actuelle qui prédit la présence d'amiante si au moins une règle conclut sur cette décision. Nous pourrions par exemple définir une technique de vote qui prendrait en compte les scores et les classes de toutes les règles applicables. Ces différentes stratégies pourraient alors être comparées à la stratégie actuelle pour sélectionner celle qui permet d'obtenir les meilleurs résultats.

7.2.2 Perspectives à long terme

• CRA-Miner n'est pas destiné à fournir des règles en temps réel, mais le temps d'exécution sur des milliers de diagnostics prend plusieurs jours. Afin d'améliorer l'efficacité de cette approche qui découvre des règles assez complexes, car riches en nombre d'atomes et dont les prédicats sont de nature variée (des prédicats contextuels, des prédicats de comparaison avec des valeurs numériques de référence), nous souhaitons explorer d'autres approches de classification dont nous pourrions nous inspirer pour optimiser CRA-Miner.

• CRA-Miner utilise un biais de langage pour contrôler l'espace de recherche qui est défini par le contexte conceptuel et les différents seuils associés. Nous avons considéré que les décisions associées à chaque produit étaient influencées par ce contexte, mais que chaque produit pouvait correspondre à un produit commercialisé différent. Or, nous pensons qu'il existe de nombreux bâtiments où le même produit commercialisé a été utilisé dans des différentes parties. Ne pas prendre en compte ce biais peut nous conduire à surestimer (ou sous-estimer) le *head coverage* et la confiance de la règle. Une perspective de notre travail serait de déterminer sur des diagnostics existants l'importance de ce biais, en vérifiant pour chaque classe la distribution des produits amiantés et non amiantés par bâtiments. Cette information pourrait être utilisée pour former une base d'apprentissage non biaisée (où l'on ne sélectionnerait pas plusieurs produits de la même classe apparaissant dans le même bâtiment ou la même structure) ou pour résoudre les doubles décisions en suivant celle de l'ensemble majoritaire afin de privilégier l'hypothèse que les mêmes produits ont été utilisés.

• Une autre perspective serait d'étudier la possibilité d'utiliser des règles de raisonnement plus complexes que celles exprimées en SWRL. Un autre langage de requête pour OWL qui est SQWRL [42] fournit un ensemble plus

large de prédicats prédéfinis qui peuvent être utilisés pour effectuer des opérations permettant des formes limitées de la négation par l'échec, le comptage et l'agrégation. L'utilité de tels opérateurs reste à étudier et à discuter avec l'expert pour trouver des cas d'usages possibles.

• CRA-Miner utilise aussi les relations partie-tout pour découvrir des règles pertinentes, mais dans un graphe de connaissance le contexte peut devenir très complexe quand on considère les composants et les relations entre les composants. Si le contexte utilisé dans une règle devient trop complexe, peu d'instances seront concernées par la règle. Aussi, nous pourrions définir une approche de classification plus flexible qui permet, au lieu de seulement décider pour les produits ayant le même contexte, de décider également pour des produits ayant un contexte assez similaire. Il s'agirait de définir une mesure de similarité reposant sur la sémantique de l'ontologie et la structure du contexte (p. ex. nombre de produits frères) pour appliquer une méthode basée sur la recherche des produits diagnostiqués les plus similaires (p. ex. méthode des plus proches voisins).

• Les deux approches ont été définies pour le problème de l'amiante. La dernière perspective à long terme serait d'étudier si les approches que nous avons définies sont applicables à d'autres domaines. Au sein du CSTB, ces approches vont être adaptées et utilisées pour la détection d'autres produits nocifs comme le plomb dans les bâtiments, ainsi que dans un projet dans le cadre de l'économie circulaire pour distinguer les éléments du bâtiment qui peuvent être recyclés de ceux qui doivent être détruits. CRA-Miner pourrait peut-être être adapté à d'autres domaines qui nécessitent de découvrir des règles de prédiction utilisant un contexte sémantique où les relations partie-tout, les relations de subsomptions et/ou les données temporelles ou numériques jouent un rôle important. Si l'on prend l'exemple du domaine de la chimie, pour prédire l'efficacité d'une solution chimique par exemple, les relations partie-tout entre les composants moléculaires de la solution peuvent être importantes pour prédire une réaction chimique. Comme certaines familles de composants chimiques sont plus efficaces que d'autres dans certains types de réactions, nous pensons que l'exploration de la sémantique qui représenterait la hiérarchie des familles de composants pourrait améliorer la prédiction de l'efficacité d'une solution dans certains types de réactions. L'utilité de l'utilisation des valeurs numériques dans les règles dans ce domaine serait de comparer, par exemple, le volume de la solution ou la quantité de certains composants dans la solution avec un certain seuil de référence calculé. Nous pensons qu'il est également possible d'utiliser des heuristiques dédiées aux relations partie-tout dans d'autres domaines comme le domaine géographique où la distribution géographique des zones géographiques peut affecter certaines propriétés de ces parties (c.-à-d. le fait qu'une zone géographique soit proche ou incluse dans une autre zone géographique peut influencer par exemple certaines caractéristiques économiques).

Annexe A

L'architecture et le fonctionnement de l'outil de prédiction de présence d'amiante

Nous avons décrit dans le chapitre 4 la procédure de construction, d'enrichissement et de peuplement de l'ontologie ASBESTOS. Nous avons montré ensuite dans les chapitres 5 et 6 le fonctionnement de l'approche hybride et de CRA-Miner. Nous décrivons dans cette annexe la communication entre le module de construction de l'ontologie et les approches de prédiction, ainsi que l'architecture et les détails de fonctionnement de chaque module. Ces modules sont assemblés dans un seul outil afin de faciliter l'utilisation par les experts d'amiante à CSTB et pour fournir une meilleure représentation de données ainsi que les résultats pour permettre une simple interprétation de ses données et résultats par les experts.

Afin de faciliter l'utilisation des deux approches de prédiction d'amiante dans les bâtiments ainsi que leur combinaison, nous avons mis en place l'outil "AsbestosReveal" qui permet de charger et traiter les données et d'exécuter les approches de prédiction.

Nous commençons d'abord par présenter le schéma général du réseau d'interaction entre les différents modules majeurs de AsbestosReveal (Section A.1). Dans cette section nous montrons aussi le fonctionnement de chaque module. Nous montrons dans la Section A.2 suivante les dépendances de l'implémentation de AsbestosReveal. La dernière Section A.3 présente des exemples de différentes règles découvertes par CRA-Miner ainsi que des explications pour chacune de ces règles.

A.1 L'architecture de AsbestosReveal

Nous représentons dans la Figure A.1 les trois modules de AsbestosReveal, Chacun de ces modules est responsable d'une fonctionnalité de AsbestosReveal.

Nous avons utilisé d'abord Protégé pour créer la partie haute de l'ontologie ASBESTOS (le premier module de l'ontologie de la Figure 4.5) qui continet les concepts majeurs et ces propriétés qui représentent le bâtiment, ses structures, ses localisations et ses produits. Cette étape initiale génère un fichier .owl de format RDF.

Le premier module contient les différentes fonctions d'enrichissement et de peuplement de l'ontologie de base (les fonctionnalités de ce module sont décrites dans le chapitre 4). La première fonction est celle responsable de l'extraction de données depuis les projets homologués qui sont initialement semi-structurés (ils sont décrits dans 4.4.1) pour créer un document structuré et utilisable par le module d'enrichissement et de peuplement de l'ontologie. Ce dernier module utilise un fichier .owl et une source de données qui sont dans ce cas l'ontologie ASBESTOS et les projets homologués structurés, pour enrichir d'abord l'ontologie avec les nouveaux concepts mentionnés dans les projets homologués ensuite peupler cette ontologie avec les individus (les données). Le résultat de ce premier module est une ontologie enrichie et peuplée avec des produits non classés (à ce point nous ne savons pas si un produit est amianté ou non).

À ce point, l'ontologie peut être utilisée par le module 2 ou le module 3 ou les deux. Le module 2 exécute l'approche hybride et utilise le module de l'ontologie représenté dans la Figure 5.1. Comme l'approche hybride (Chapitre 5) a besoin de ressources externes pour calculer les probabilités et de diagnostics pour trouver le meilleur seuil, nous utilisons d'abord le module d'enrichissement et de peuplement de l'ontologie pour intégrer ces ressources avant d'appeler la fonction de l'approche hybride qui utilise les concepts de la Figure 5.1 pour associer les produits à des classes (positives ou négatives). Le module 3 de AsbestosReveal utilise aussi le module 3 de l'ontologie (Figure 6.1) afin de représenter les résultats de CRA-Miner (Chapitre 6). CRA-Miner utilise le module d'enrichissement et de peuplement de l'ontologie pour ajouter les diagnostics à l'ontologie puis il appelle la fonction de CRA-Miner qui utilise les diagnostics pour trouver des règles de raisonnement. Ensuite ces règles vont être utilisées dans l'étape de raisonnement avec Pellet pour classifier les produits.

À la fin de la procédure, nous obtenons l'ontologie ASBESTOS avec les produits de projets homologués classés, les décisions associées aux produits peuvent venir de l'approche hybride ou de CRA-Miner. Ces décisions seront observées et examinées par l'expert de CSTB.

A.1.1 Le module d'extraction de données

La Figure A.2 illustre les détails du processus d'extraction des données à partir de différents types de documents. Dans le cas des diagnostics et des ressources externes, comme les données sont représentées sous format tabulaire où chaque terme est classé dans une colonne qui représente son type (son concept), l'extracteur suit le modèle utilisé dans chaque document (l'ordre de colonnes et les types d'informations dans chaque document). Dans le cas des projets homologués, qui suivent aussi une structuration tabulaire, des descriptions textuelles sont utilisées pour décrire les localisations et les produits (comme montré dans la Section 4.8.2) au lieu d'une représentation sous

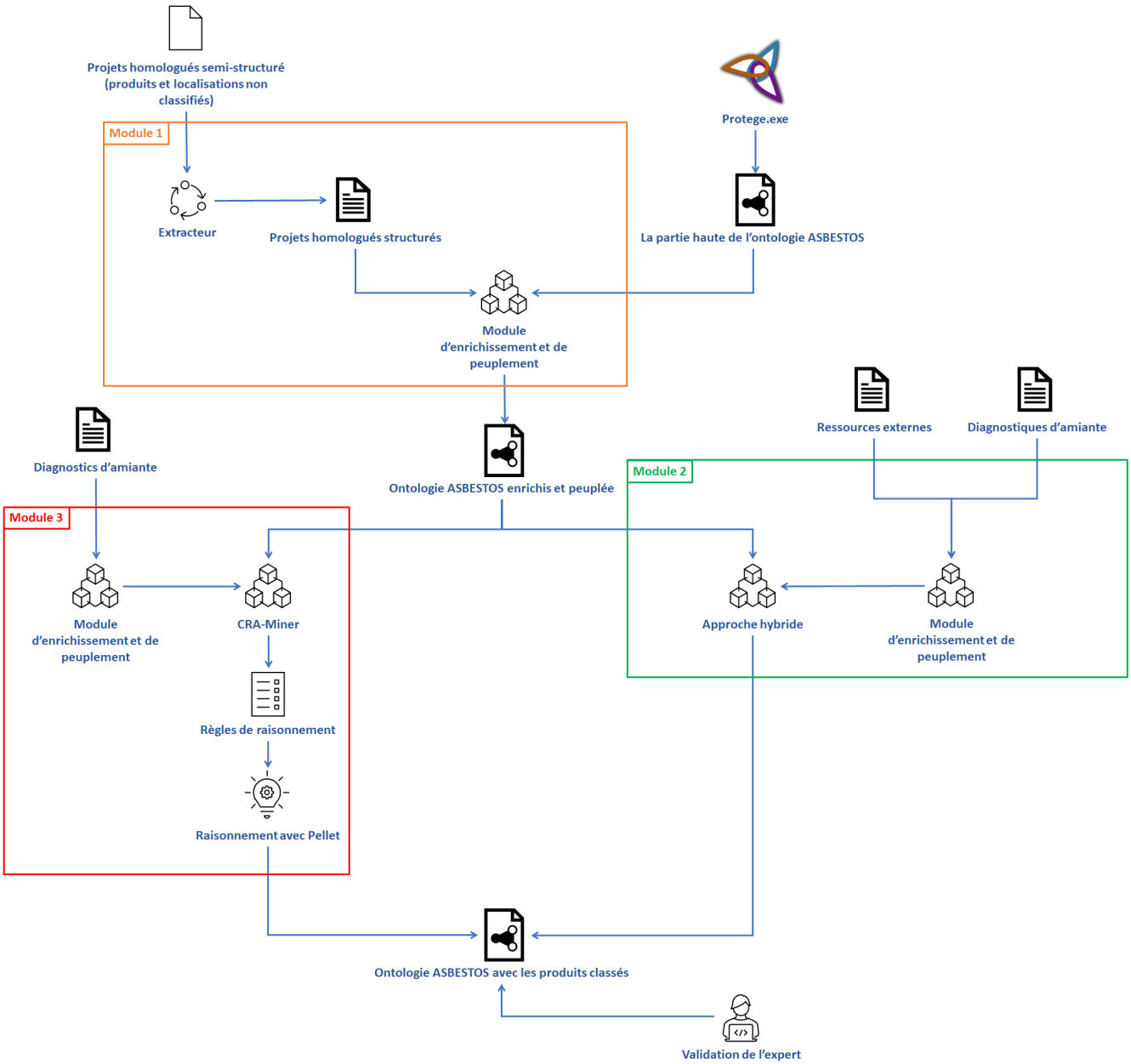


FIGURE A.1 – Les modules de AsbestosReveal

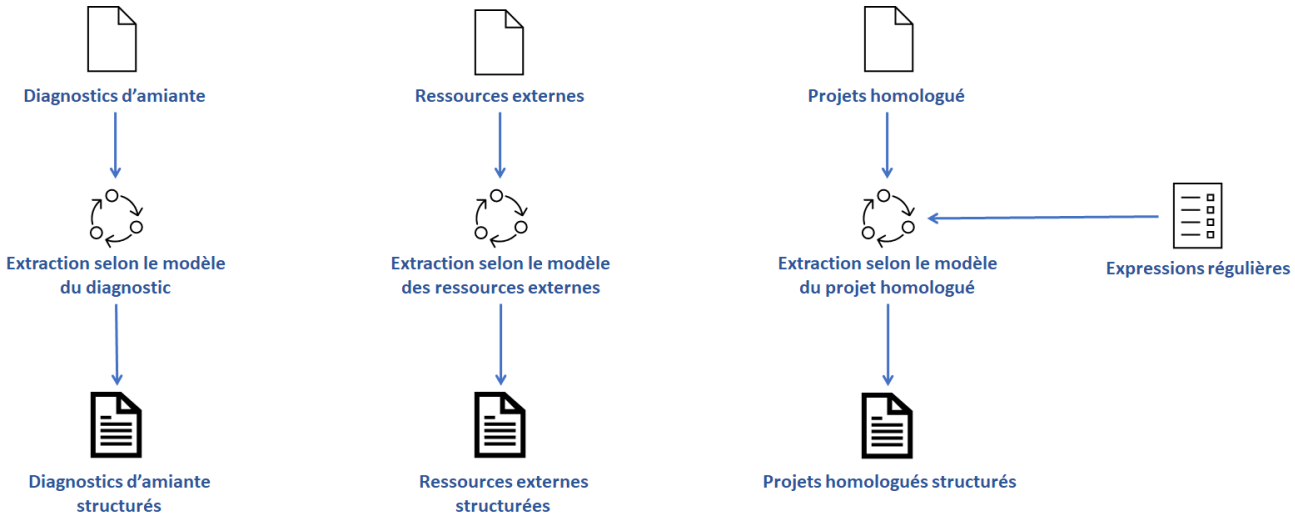


FIGURE A.2 – Le processus d'extraction selon le type du document

forme de colonnes séparées. Afin d'extraire les termes de localisations et des produits depuis ces descriptions textuelles, l'extracteur utilise une expression régulière pour automatiquement reconnaître les termes et ensuite les classifier selon leur positionnement dans le texte. Le résultat de l'extraction est un document tabulaire CSV qui peut être parsé par le module d'enrichissement et de peuplement.

A.1.2 Le module d'enrichissement et de peuplement de l'ontologie

Les documents formalisés qui résultent de la procédure d'extraction seront utilisés par le module d'enrichissement et de peuplement dans le but d'ajouter les nouvelles données, ainsi que les nouveaux concepts, à la version de l'ontologie ASBESTOS actuelle. La Figure A.3 montre la combinaison des trois modules de l'ontologie de base. Dans le cas des projets homologués, le module d'enrichissement et de peuplement ajoute d'abord les nouveaux concepts. Ensuite il instancie les concepts pour créer les individus mentionnés dans les données qui représentent les composants d'un bâtiment et ces propriétés (l'année, le type, etc.). Dans le cas des diagnostics, le module d'enrichissement et de peuplement instancie, en plus des composants, le concept "Diagnostic characteristic" afin de représenter les données diagnostiquées par le diagnostic 0 qui signifie l'absence d'amiante ou par 1 qui signifie la présence d'amiante. Le cas des données qui proviennent de ressources externes est aussi différent, car ces dernières possèdent des caractéristiques supplémentaires pour décrire des produits commercialisés. Dans ce cas, le module d'enrichissement et de peuplement ajoute d'abord le nom du produit commercialisé, le nom de son fournisseur et le type d'amiante qu'il possède (si ces informations son mentionnées) dans les propriétés du produit. Ensuite, il instancie le concept "Extracted characteristic" dans le but de représenter les données temporelles. Ces données temporelles sont décrites par un intervalle de temps avec une année de début et une année de fin, la

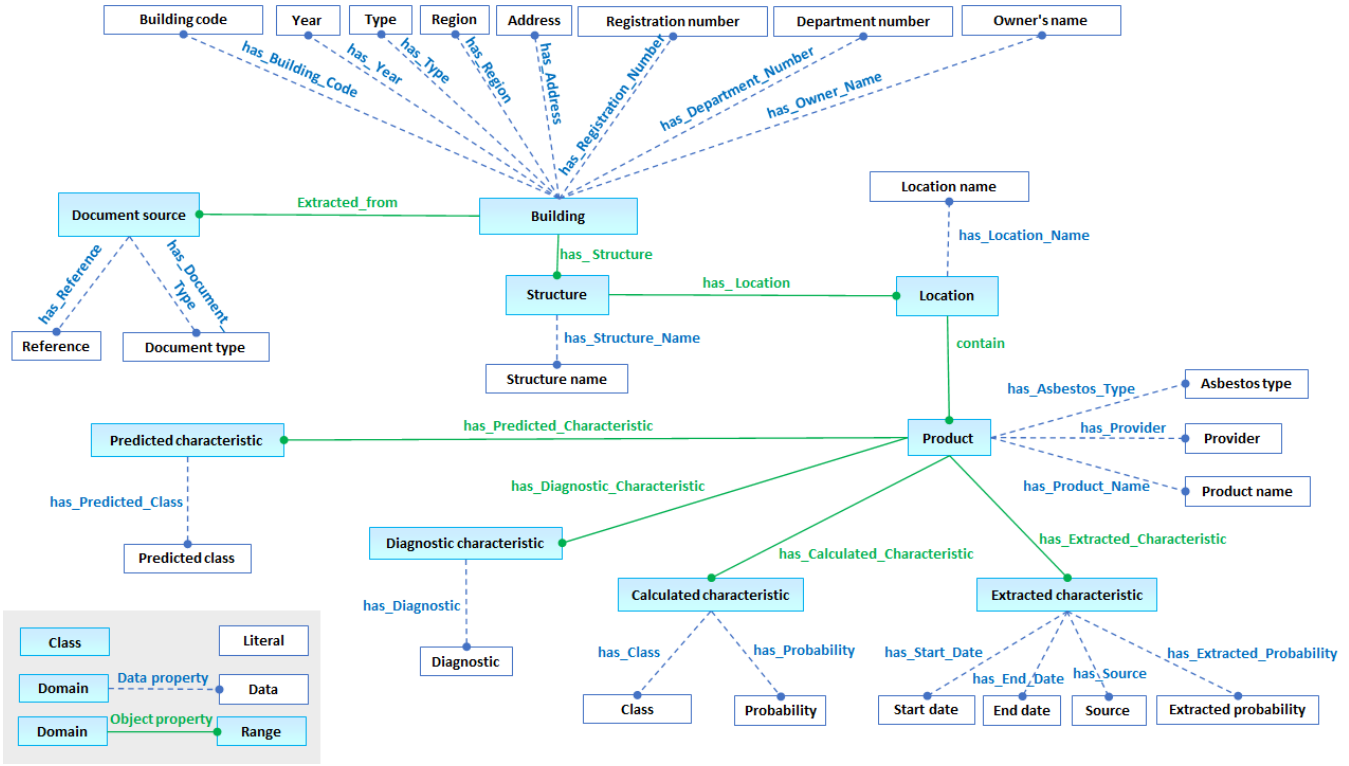


FIGURE A.3 – L'ontologie de base complète (avec les trois modules intégrés)

source de la donnée (INRS, ANDEVA ou FUSION) et la probabilité de présence d'amiante (0 si le produit est non amianté, 1 s'il est amianté ou $0,5$ si l'information est inconnue). La sortie du module d'enrichissement et de peuplement est l'ontologie ASBESTOS enrichie est peuplée avec les données des documents à l'entrée du module. La Figure A.4 montre l'ensemble des concepts créés par ce module affichés dans Protégé. L'ontologie ASBESTOS contient 139 concepts, dont 8 sont des sous-concepts de "Structure", 20 sont des sous-concepts de "Localisation" et 102 sont des sous-concepts de "Produit".

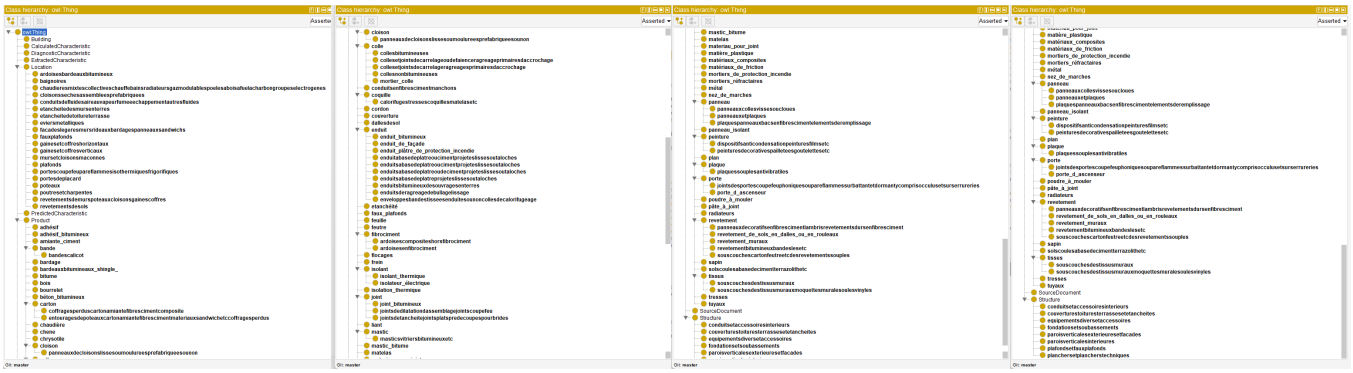


FIGURE A.4 – Extrait de concepts de Protégé

A.2 Les dépendances de AsbestosReveal et les approches de prédiction

Nous avons implémenté AsbestosReveal ainsi que CRA-Miner et l'approche hybride en utilisant Python 3. Nous avons aussi utilisé un ensemble de bibliothèques pour réaliser certaines fonctionnalités de AsbestosReveal :

— owlready2 [1] : afin de manipuler les ontologies (lire, modifier,etc.), nous avons utilisé le package owlready2 qui est utilisé pour la programmation orientée ontologie en Python. Il peut charger des ontologies OWL 2.0 en tant qu'objets Python, les modifier, les enregistrer et effectuer un raisonnement via HermiT ou Pellet (inclus).

— rdflib [2] : cette bibliothèque permet la manipulation des graphes RDF. Elle offre aussi la possibilité de charger les fichiers owl comme des graphes RDF.

— treetaggerwrapper [3] : cette bibliothèque contient un système d'étiquetage de partie de discours en permettant d'étiqueter les mots dans un passage textuel avec ses natures (noms, verbes, adjectifs, etc.). Cette bibliothèque est utilisée par le module d'extraction de AsbestosReveal afin de détecter les termes.

A.3 Les règles de raisonnement apprises par CRA-Miner

Lors de son exécution, CRA-Miner apprend d'abord un ensemble de règles depuis les diagnostics amiante qu'il utilise par la suite pour raisonner avec Pellet. Nous décrivons dans cette section quelques exemples de règles trouvées par CRA-Miner (les données ainsi que l'ensemble complet de règles sont disponibles dans le GitHub [4]).

La règle suivante :

bandescalicot(?P), has_diagnostic_characteristc(?P, ?D) $\rightarrow$ *has_diagnostic(?D, "positif")*

est une règle qui englobe tous les produits de type "bandescalicot" et les classe comme des positifs. Cette règle a une confiance de 1 ce qui signifie qu'elle est applicable pour tous les produits de type "bandescalicot". Le *head coverage* de cette règle $hc = 0,03$ signifie que les produits de type "bandescalicot" représentent 3% de l'ensemble de produits positifs.

La règle :

enduitsabasedeplatreoudecimentprojeteslissesoutaloches(?P), has_location(?S, ?L), contain(?L, ?P), has_structure(?B, ?S), has_year(?B, ?Y), has_diagnostic_characteristic(?P, ?D) , greaterThanOrEqual(?Y, "1988-01-01T00 :00 :00") $\rightarrow$ *has_diagnostic(?D, "negatif")*

montre un exemple des règles qui explorent une contrainte temporelle qui est l'année de construction du bâtiment (CRA-Miner a trouvé 14 règles en moyenne qui exploitent une contrainte temporelle). CRA-Miner a trouvé que

1. Documentation d'Owlready2 : `https://owlready2.readthedocs.io/en/v0.36/`
2. rdflib 6.1.1 : https://rdflib.readthedocs.io/en/stable/
3. Documentation de TreeTagger Python Wrapper : https://treetaggerwrapper.readthedocs.io/en/latest/
4. Le répertoire GitHub des données utilisées dans les tests ainsi que les règles découvertes : `https://github.com/ThamerMECHARNIA/DATA-IC2021`

l'année de référence qui maximise la confiance de la règle est 1988. La confiance de cette règle est $= 1$ et son $hc = 0,4$ ce qui signifie que les produits de type "enduitsabasedeplatreoudecimentprojeteslissesoutaloches" qui sont utilisés dans des bâtiments construits après 1988 représentent 4% des produits négatifs.

CRA-Miner a trouvé aussi 29 règles qui impliquent 1 à 2 produits frères et 17 règles qui impliquent 1 à 3 localisations frères. Les deux exemples suivants montrent cette exploration de contexte par CRA-Miner :

collesetjointsdecarrelageragreagesprimairesdaccrochage(?P), has_location(?S, ?L), contain(?L, ?P), souscouchescartonfeutreetcdesrevetementssouples(?P3), contain(?L, ?P3), dallesdesol(?P2), contain(?L, ?P2), has_diagnostic_characteristic(?P, ?D) $\rightarrow$ *has_diagnostic(?D, "positif")*

cette règle contient deux produit frères (P2 de type "dallesdesol" et P3 de type "souscouchescartonfeutreetcdesrevetementssouples") co-localisé dans la même localisation que le produit principal de la règle qui est de type "collesetjointsdecarrelageragreagesprimairesdaccrochage". Cette règle a une confiance de $= 1$ et a un $hc = 0,006$.

La règle suivante :

enduitsabasedeplatreprojeteslissesoutaloches(?P), has_location(?S, ?L), poteaux(?L3), has_location(?S, ?L3), revetementsdemurspoteauxcloisonsgainescoffres(?L2), has_location(?S, ?L2), contain(?L, ?P), has_diagnostic_characteristc(?P, ?D) $\rightarrow$ *has_diagnostic(?D, "positif")*

est un exemple d'exploration de deux localisations frères (L2 de type "revetementsdemurspoteauxcloisonsgainescoffres" et L3 de type "poteaux"). Cette règle a un $hc = 0,01$ et une confiance de $= 0,6$.

Bibliographie

[1] H. Adé, L. De Raedt, and M. Bruynooghe. Declarative bias for specific-to-general ilp systems. *Machine Learning*, 20(1) :119–154, 1995.

[2] H. Blockeel and L. De Raedt. Top-down induction of first-order logical decision trees. *Artificial intelligence*, 101 (1-2) :285–297, 1998.

[3] H. Blockeel, L. Dehaspe, J. Ramon, J. Struyf, A. Van Assche, C. Vens, and D. Fierens. The ace data mining system, user's manual. *Katholieke Universiteit Leuven, Belgium*, 2006.

[4] H. Boström. Covering vs. divide-and-conquer for top-down induction of logic programs. In *IJCAI*, pages 1194–1200, 1995.

[5] H. Boström and L. Asker. Combining divide-and-conquer and separate-and-conquer for efficient and effective rule induction. In *International Conference on Inductive Logic Programming*, pages 33–43. Springer, 1999.

[6] Y. Bouchard. Le modèle tout-partie dans l'ontologie de louis lavelle. *Revue de Métaphysique et de Morale*, pages 351–378, 1999.

[7] A. Cropper and S. Tourret. Logical reduction of metarules. *Machine Learning*, 109(7) :1323–1369, 2020.

[8] C. d'Amato, A. G. B. Tettamanzi, and D. M. Tran. Evolutionary discovery of multi-relational association rules from ontological knowledge bases. In E. Blomqvist, P. Ciancarini, F. Poggi, and F. Vitali, editors, *EKAW 2016, Italy, November 19-23, 2016, Proceedings*, volume 10024 of *Lecture Notes in Computer Science*, pages 113–128, 2016.

[9] H. C. de la Santé Publique et al. Repérage de l'amiante, mesures d'empoussièrement et révision du seuil de déclenchement des travaux de retrait ou de confinement de matériaux contenant de l'amiante–analyse et recommandations [validé par la commission spécialisée «risques liés à l'environnement» le 23 mai 2014]. *Paris, Haut Conseil de la Santé Publique*, 2014.

[10] L. De Raedt. Logical settings for concept-learning. *Artificial Intelligence*, 95(1) :187–201, 1997.

[11] S. Desprès. Construction d'une ontologie modulaire. application au domaine de la cuisine numérique. *Rev. d'Intelligence Artif.*, 30(5) :509–532, 2016.

[12] Y. Ding and S. Foo. Ontology research and development. part 1-a review of ontology generation. *Journal of information science*, 28(2) :123–136, 2002.

[13] J. Euzenat, P. Shvaiko, et al. *Ontology matching*, volume 18. Springer, 2007.

[14] N. Fanizzi, C. d'Amato, and F. Esposito. DL-FOIL concept learning in description logics. In *Inductive Logic Programming, ILP 2008, Czech Republic, September 10-12, 2008, Proceedings*, volume 5194 of *Lecture Notes in Computer Science*, pages 107–121, 2008.

[15] M. Fernández-López. Overview of methodologies for building ontologies. In *IJCAI99 workshop on ontologies and problem-solving methods : Lessons learned and future trends*, volume 430. Citeseer, 1999.

[16] M. Fernández-López, A. Gómez-Pérez, and N. Juristo. Methontology : from ontological art towards ontological engineering. 1997.

[17] N. Fornara, D. Okouya, and M. Colombetti. Using owl 2 dl for expressing acl content and semantics. In *European workshop on multi-agent systems*, pages 97–113. Springer, 2011.

[18] J. Fürnkranz. Separate-and-conquer rule learning. *Artificial Intelligence Review*, 13(1) :3–54, 1999.

[19] J. Fürnkranz and T. Kliegr. A brief overview of rule learning. In N. Bassiliades, G. Gottlob, F. Sadri, A. Paschke, and D. Roman, editors, *Rule Technologies : Foundations, Tools, and Applications - 9th International Symposium, RuleML 2015, Berlin, Germany, August 2-5, 2015, Proceedings*, volume 9202 of *Lecture Notes in Computer Science*, pages 54–69. Springer, 2015.

[20] L. Galárraga, C. Teflioudi, K. Hose, and F. M. Suchanek. Fast rule mining in ontological knowledge bases with AMIE+. *VLDB Journal*, 24(6) :707–730, 2015. doi : 10.1007/s00778-015-0394-1. URL `https://doi.org/10.1007/s00778-015-0394-1`.

[21] A. Gómez-Pérez and M. C. Suárez-Figueroa. Neon methodology for building ontology networks : a scenario-based methodology. 2009.

[22] A. Grilo and R. Jardim-Goncalves. Value proposition on interoperability of bim and collaborative working environments. *Automation in construction*, 19(5) :522–530, 2010.

[23] T. R. Gruber. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2) :199–220, 1993.

[24] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, M. Dean, et al. Swrl : A semantic web rule language combining owl and ruleml. *W3C Member submission*, 21(79) :1–31, 2004.

[25] T. Inoue, K. Takano, T. Watanabe, J. Kawahara, R. Yoshinaka, A. Kishimoto, K. Tsuda, S.-i. Minato, and Y. Hayashi. Distribution loss minimization with guaranteed error bound. *IEEE Transactions on Smart Grid*, 5(1) : 102–111, 2014.

[26] J.-U. Kietz, R. Volz, and A. Maedche. Extracting a domain-specific ontology from a corporate intranet. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*, 2000.

[27] J. Lajus, L. Galárraga, and F. Suchanek. Fast and exact rule mining with amie 3. In *Extended Semantic Web Conference (ESWC)*, volume 12123 of *Lecture Notes in Computer Science*, pages 36–52. Springer, 2020.

[28] O. Lassila and D. McGuinness. The role of frame-based representation on the semantic web. *Linköping Electronic Articles in Computer and Information Science*, 6(5) :2001, 2001.

[29] C.-H. Lee. Towards implementing robust visual motion computations : Generate and test approach. 1988.

[30] M. F. López, A. Gómez-Pérez, J. P. Sierra, and A. P. Sierra. Building a chemical ontology using methontology and the ontology design environment. *IEEE Intelligent Systems and their applications*, 14(1) :37–46, 1999.

[31] A. Maedche and S. Staab. Mining ontologies from text. In *International conference on knowledge engineering and knowledge management*, pages 189–202. Springer, 2000.

[32] T. Mecharnia, L. C. Khelifa, N. Pernelle, and F. Hamdi. An approach toward a prediction of the presence of asbestos in buildings based on incomplete temporal descriptions of marketed products. In M. Kejriwal, P. A. Szekely, and R. Troncy, editors, *K-CAP 2019, Marina Del Rey, CA, USA, November 19-21, 2019*, pages 239–242. ACM, 2019.

[33] T. Mecharnia, N. Pernelle, L. C. Khelifa, and F. Hamdi. Approche de prédiction de présence d'amiante dans les bâtiments basée sur l'exploitation des descriptions temporelles incomplètes de produits commercialisés. 2019.

[34] T. Mecharnia, L. C. Khelifa, F. Hamdi, N. Pernelle, and C. Rouveirol. Découverte de règles contextuelles pour prédire la présence d'amiante dans les bâtiments. In *Journées Francophones d'Ingénierie des Connaissances (IC) Plate-Forme Intelligence Artificielle (PFIA'21)*, pages pp–73, 2021.

[35] T. Mecharnia, N. Pernelle, C. Rouveirol, F. Hamdi, and L. C. Khelifa. Mining contextual rules to predict asbestos in buildings. In *International Conference on Conceptual Structures*, pages 170–184. Springer, 2021.

[36] T. M. Mitchell. Generalization as search. *Artificial intelligence*, 18(2) :203–226, 1982.

[37] S. Muggleton. Inverse entailment and progol. *New generation computing*, 13(3) :245–286, 1995.

[38] S. Muggleton and L. De Raedt. Inductive logic programming : Theory and methods. *The Journal of Logic Programming*, 19 :629–679, 1994.

[39] S. Muggleton and L. D. Raedt. Inductive logic programming : Theory and methods. *J. Log. Program.*, 19/20 : 629–679, 1994. doi : $10.1016/0743\text{-}1066(94)90035\text{-}3$. URL `https://doi.org/10.1016/0743-1066(94)90035-3`.

[40] S. Muggleton, C. Feng, et al. *Efficient induction of logic programs*. Citeseer, 1990.

[41] A. K. Nassiri, N. Pernelle, F. Saïs, and G. Quercini. Generating referring expressions from rdf knowledge graphs for data linking. In *International Semantic Web Conference*, pages 311–329. Springer, 2020.

[42] M. J. O'Connor and A. K. Das. Sqwrl : a query language for owl. In *OWLED*, volume 529, 2009.

[43] S. Ortona, V. V. Meduri, and P. Papotti. Robust discovery of positive and negative rules in knowledge bases. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 1168–1179, 2018.

[44] H. Paulheim, V. Tresp, and Z. Liu. Representation learning for the semantic web. *J. Web Semant.*, 61-62 : 100570, 2020.

[45] H. S. Pinto, S. Staab, and C. Tempich. Diligent : Towards a fine-grained methodology for distributed, loosely-controlled and evolving engineering of ontologies. In *ECAI*, volume 16, page 393. Citeseer, 2004.

[46] J. R. Quinlan. Induction of decision trees. *Machine learning*, 1(1) :81–106, 1986.

[47] J. R. Quinlan. Learning logical definitions from relations. *Mach. Learn.*, 5 :239–266, 1990.

[48] G. Rizzo, N. Fanizzi, and C. d'Amato. Class expression induction as concept space exploration : From dl-foil to dl-focl. *Future Gener. Comput. Syst.*, 108 :256–272, 2020.

[49] I. J. Selikoff, D. H. K. Lee, et al. *Asbestos and disease.* Academic Press, Inc. 111 Fifth Avenue, New York, New York 10003, USA, 1978.

[50] A. Srinivasan. The aleph manual, 2001.

[51] S. Staab, R. Studer, H.-P. Schnurr, and Y. Sure. Knowledge processes and ontologies. *IEEE Intelligent systems*, 16(1) :26–34, 2001.

[52] D. Symeonidou, V. Armant, N. Pernelle, and F. Saïs. Sakey : Scalable almost key discovery in rdf data. In *International Semantic Web Conference*, pages 33–49. Springer, 2014.

[53] D. Symeonidou, L. Galárraga, N. Pernelle, F. Saïs, and F. Suchanek. Vickey : Mining conditional keys on knowledge bases. In *International Semantic Web Conference*, pages 661–677. Springer, 2017.

[54] N. N. Than and I. Baimuratov. Logic graphs for alc, shif and shoin description logics. In *Conference of Open Innovations Association, FRUCT*, number 27, pages 386–389. FRUCT Oy, 2020.

[55] R. Virta. Asbestos. *Kirk-Othmer Encyclopedia of Chemical Technology*, pages 1–40, 2000.

[56] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Silk-a link discovery framework for the web of data. In *Ldow*, 2009.

[57] A. Wagner. Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis. In *ECAI Workshop on Ontology Learning*, volume 61. Citeseer, 2000.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Approches sémantiques pour la prédiction de présence d'amiante dans les bâtiments : une approche probabiliste et une approche à base de règles

Mots clés : Fouille de règles - Ontologie - Graphe de connaissances - Données temporelles - Incomplétude des données - Amiante.

Résumé : De nos jours, les Graphes de Connaissances sont utilisés pour représenter toutes sortes de données et ils constituent des ressources évolutives, interopérables et exploitables par des outils d'aide à la décision. Le Centre Scientifique et Technique du Bâtiment (CSTB) a été sollicité pour développer un outil d'aide à l'identification des matériaux contenant de l'amiante dans les bâtiments. Dans ce contexte, nous avons créé et peuplé l'ontologie ASBESTOS qui permet la représentation des données des bâtiments et les résultats des diagnostics réalisés en vue de détecter la présence d'amiante dans les produits utilisés. Nous nous sommes ensuite basés sur ce graphe de connaissance pour développer deux approches qui permettent de prédire la présence d'amiante dans les produits en l'absence de la référence du produit commercialisé effectivement utilisé.
La première approche, nommée approche hybride, se base sur des ressources externes décrivant les périodes où les produits commercialisés sont amiantés pour calculer une probabilité d'existence d'amiante dans un composant du bâtiment. Cette approche traite les conflits entre les ressources externes, et l'incomplétude des données répertoriées en appliquant une approche de fusion pessimiste qui ajuste les probabilités calculées en utilisant un sous-ensemble de diagnostics.
La deuxième approche, nommée CRA-Miner, s'inspire de méthodes de programmation logique inductive (PLI) pour découvrir des règles à partir du graphe de connaissances décrivant les bâtiments et les diagnostics d'amiante. La référence des produits spécifiques utilisés lors de la construction n'étant jamais spécifiée, CRA-Miner considère les données temporelles, la sémantique de l'ontologie ASBESTOS, les types de produits et les informations contextuelles telles que les relations partie-tout pour découvrir un ensemble de règles qui pourront être utilisées pour prédire la présence d'amiante dans les éléments de construction.
L'évaluation des deux approches menées sur l'ontologie ASBESTOS peuplée avec les données fournies par le CSTB montre que les résultats obtenus, en particulier quand les deux approches sont combinées, sont tout à fait prometteurs.

Title : Semantic approaches for predicting the presence of asbestos in buildings : a probabilistic approach and a rule-based approach

Keywords : Rule mining - Ontology - Knowledge graph - Temporal data - Data incompleteness - Asbestos.

Abstract : Nowadays, Knowledge Graphs are used to represent all kinds of data and they constitute scalable and interoperable resources that can be used by decision support tools. The Scientific and Technical Center for Building (CSTB) was asked to develop a tool to help identify materials containing asbestos in buildings. In this context, we have created and populated the ASBESTOS ontology which allows the representation of building data and the results of diagnostics carried out in order to detect the presence of asbestos in the used products. We then relied on this knowledge graph to develop two approaches which make it possible to predict the presence of asbestos in products in the absence of the reference of the marketed product actually used.
The first approach, called the hybrid approach, is based on external resources describing the periods when the marketed products are asbestos-containing to calculate the probability of the existence of asbestos in a building component. This approach addresses conflicts between external resources, and incompleteness of listed data by applying a pessimistic fusion approach that adjusts the calculated probabilities using a subset of diagnostics.
The second approach, called CRA-Miner, is inspired by inductive logic programming (ILP) methods to discover rules from the knowledge graph describing buildings and asbestos diagnoses. Since the reference of specific products used during construction is never specified, CRA-Miner considers temporal data, ASBESTOS ontology semantics, product types and contextual information such as part-of relations to discover a set of rules that can be used to predict the presence of asbestos in construction elements.
The evaluation of the two approaches carried out on the ASBESTOS ontology populated with the data provided by the CSTB show that the results obtained, in particular when the two approaches are combined, are quite promising.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France