Approche de prédiction de présence d'amiante dans les bâtiments basée sur l'exploitation des descriptions temporelles incomplètes de produits commercialisés

Thamer Mecharnia, Nathalie Pernelle, Lydia Chibout Khelifa, Fayçal Hamdi

Approche de prédiction de présence d'amiante dans les bâtiments basée sur l'exploitation des descriptions temporelles incomplètes de produits commercialisés

Thamer Mecharnia[1,2], Nathalie Pernelle[1], Lydia Chibout Khelifa[2], Fayçal Hamdi[3]

[1] LRI, Université Paris sud, Orsay, France
prenom.nom@lri.fr

[2] CENTRE SCIENTIFIQUE ET TECHNIQUE DU BÂTIMENT (CSTB), Champs sur Marne, France
prenom.nom@cstb.fr

[3] CEDRIC – CNAM, Paris, France
faycal.hamdi@cnam.fr

Résumé : La production, l'importation et la commercialisation d'amiante sont interdites depuis le premier janvier 1997 en France. Cependant, il en reste des millions de tonnes disséminés dans les usines, les immeubles, les établissements scolaires, ou encore les hôpitaux.

Dans cet article nous proposons une méthode de prédiction de présence d'amiante basée sur des données temporelles décrivant la probabilité de présence d'amiante dans des produits commercialisés pour calculer une probabilité d'existence de produits amiantés dans les bâtiments. Pour atteindre notre but, nous proposons une ontologie amiante qui va être peuplée en utilisant les données issues de ressources externes. Ensuite, ces informations sont utilisées pour calculer la probabilité d'amiante pour les éléments constituant un bâtiment donné. Notre approche a été expérimentée sur des données synthétiques décrivant 120 bâtiments en s'appuyant sur les 704 produits amiantés de l'INRS et l'ANDEVA.

Mots-clés : Ontologies, Informations incertaines, données temporelles, prédiction.

1 Introduction

Pour ses qualités ignifuges, la France a abondamment utilisé l'amiante, de l'avant-guerre à son interdiction en 1997, plutôt tardive en Europe, en particulier au cours des décennies 1950 à 1970. La nocivité de l'amiante ou de l'asbeste est connue de longue date, mais son danger est identifié depuis le début du XXe siècle. Une accumulation de données scientifiques et médicales sur l'amiante telle que rapportée dans l'expertise collective de l'INSERM de 1997, rappelle les principales étapes des connaissances médicales relatives aux effets des expositions à l'amiante sur la santé.

La production, l'importation et la commercialisation d'amiante sont interdites depuis le premier janvier 1997 en France. Cependant, il en reste des millions de tonnes disséminés dans les usines, les immeubles, les établissements scolaires, ou encore les hôpitaux. 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 (e.g. fenêtre, plancher, porte, ...) dans le cadre de l'économie circulaire. Dans le cadre des travaux du Plan de recherche et développement Amiante (PRDA), le CSTB a été sollicité pour élaborer un outil en ligne fournissant une aide au repérage de matériaux amiantés dans les bâtiments (ORIGAMI) qui a pour objectif d'orienter l'opérateur de repérage et de l'aider dans la préparation de son programme de repérage. Il 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 X 46-020. L'outil pourra également être un support de formation, voire être étendu à un usage « Particuliers » dans un objectif de sensibilisation au risque amiante.

Ce projet a fait émerger de nouvelles problématiques à savoir, la pérennisation des connaissances dans le domaine de l'amiante, le partage et la réutilisation de ces connaissances

dans d'autres domaines tels que le réemploi et le recyclage des produits/matériaux pour la construction (comme le béton dans le cadre de l'économie circulaire). Répondre à ces problématiques devrait apporter une aide dans le domaine de l'emploi ou de la restriction des matériaux/produits (à valider par les Groupes de Travail Spécialisés du CSTB) ainsi que la détection de cas particuliers.
Dans cet article nous proposons une méthode d'analyse de données temporelles incertaines qui permet de calculer une probabilité d'existence de produits amiantés dans les bâtiments à partir d'un ensemble de descriptions de bâtiments et de ressources externes décrivant des produits ayant été amiantés ou probablement amiantés à certaines périodes. Pour atteindre notre but, nous proposons une ontologie amiante qui est enrichie et peuplée en utilisant les données d'entrée. Ensuite, cette ontologie peuplée est utilisée pour le calcul de la probabilité de présence d'amiante dans un élément constitutif d'un bâtiment. Notre approche a été expérimentée sur des données synthétiques décrivant 120 bâtiments en s'appuyant sur les 704 produits amiantés décrits par l'INRS et l'ANDEVA.

Dans la section 2, nous présentons les ressources disponibles au sein du CSTB et la problématique de prédiction que nous avons définie en collaboration avec les experts. Dans la section 3, nous décrivons l'ontologie amiante proposée. Puis, en section 4, nous présentons l'approche que nous avons définie pour enrichir et peupler l'ontologie par les produits commercialisés décrits dans l'INRS et l'ANDEVA et pour calculer la probabilité de l'existence de l'amiante dans une partie du bâtiment. Finalement, en section 5, nous présentons les résultats obtenus dans nos premières expérimentations.

2 Contexte et Problématique

Dans le problème que nous traitons, l'objectif est de prédire la présence potentielle d'amiante dans un bâtiment. Le CSTB archive un ensemble de documents qui décrivent les bâtiments construits en France, dont les deux types de documents décrits ci-dessous :

— **Projet type homologué** : document qui contient un ensemble d'informations décrivant un bâtiment (nom, adresse, type, année de construction, région), et la liste de ses structures (ouverture extérieure, balcon, ...). Pour chaque structure, le document décrit l'ensemble de ses localisations (porte, fenêtre, ...) ainsi que les familles de produits utilisées pour chaque localisation (enduit, colle, ...).

— **Diagnostic amiante** : document qui décrit les résultats des prélèvements effectués sur un bâtiment pour détecter la présence d'amiante dans des éléments constituant des parties de bâtiments (produits, localisations).

A l'heure actuelle, pour prédire une présence éventuelle d'amiante et demander un prélèvement, l'expert utilise les diagnostics effectués pour d'autres projets homologués. Si le bâtiment possède les mêmes caractéristiques que celui qui est mentionné dans un diagnostic (même région et même type), et si une classe de produit contient de l'amiante, il suppose que la même classe de produit peut contenir de l'amiante et il demande le prélèvement et l'analyse d'un échantillon. Dans les cas où il ne trouve pas de diagnostic qui concerne un bâtiment similaire au bâtiment en cours d'étude, il demande que des échantillons soient analysés en laboratoire pour toutes les parties du bâtiment.

L'objectif du projet est d'aider l'expert à prédire la présence de l'amiante dans les familles de produits dans un bâtiment donné en s'appuyant sur les ressources documentaires du CSTB qui couvre la période de 1943 à 1997. Comme les projets types homologués ne mentionnent pas les produits commercialisés réellement utilisés lors de la construction du bâtiment, nous faisons l'hypothèse que nous pouvons calculer une probabilité d'existence d'amiante pour un produit utilisé à partir des produits de cette famille commercialisés au moment de la construction de ce bâtiment, et que les familles de produits n'utilisent plus d'amiante conformément à l'interdiction de son utilisation à partir de 1997. Plus précisément, les différentes étapes du projet sont les suivantes :

(1) Construire une ontologie Amiante qui permette de modéliser les connaissances sur les bâtiments et les diagnostics réalisés sur les bâtiments quand ils existent.

(2) Enrichir et peupler l'ontologie par des classes et des instances de classes et de propriétés issues d'un processus d'extraction automatique des informations décrites dans les projets homologués.

(3) Enrichir l'ontologie en s'appuyant sur des ressources externes décrivant les produits commercialisés et la probabilité de présence d'amiante dans ces produits.

(4) Proposer une approche de prédiction de présence d'amiante basée sur ces connaissances.

Nous disposons pour l'instant d'un nombre insuffisant de projets homologués et de diagnostics pour définir une méthode générique pour l'étape (2) d'extraction automatique, ni pour définir une méthode supervisée qui permette d'apprendre à prédire la présence d'amiante à partir des bâtiments, des diagnostics associés et des produits commercialisés pour l'étape (4). Aussi, dans cet article, nous nous focalisons sur les étapes (1), (3) et sur la définition d'une approche non supervisée pour l'étape (4).

3 L'Ontologie Amiante

Nous avons manuellement construit la partie haute de l'ontologie Amiante en nous basant sur les ressources documentaires du CSTB, les besoins en termes de prédiction (cf. figure 1) et en interagissant avec l'expert. Les principaux concepts de cette ontologie sont les suivants :

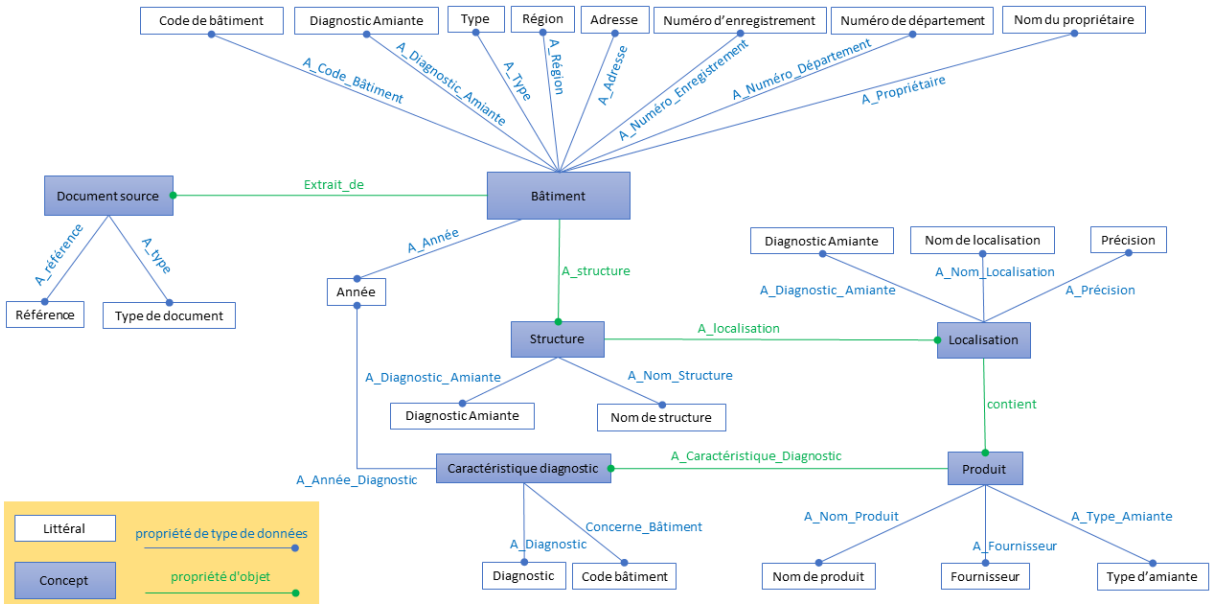

FIGURE *1 – Concepts principaux de l'ontologie Amiante*

— Bâtiment : construction qui est caractérisée par un code, un type (codification CSTB qui correspond à un genre donné de bâtiment : école, logement), une année de construction, une adresse, et la région où le bâtiment a été construit.
— Structure : composant essentiel d'un bâtiment qui correspond à un sous-espace du bâtiment (ex : terrassement, balcon, escalier, toiture, plancher-sol, ...)
— Localisation : désigne un élément de base qui appartient à une structure de bâtiment (ex : porte, fenêtre, plancher, mur, ...).
— Produit : décrit un produit qui peut entrer dans la composition des localisations (ex : colle, enduit, . . .). Un produit est décrit par son nom, le nom de fournisseur et le type d'amiante s'il est amianté.

— Document source : référence des documents, projet type homologué ou diagnostic Amiante, qui décrivent respectivement les propriétés des Bâtiments, les familles de produits ou les résultats des prélèvements effectués sur le bâtiment. Un document source est caractérisé par son type (projet type homologué ou diagnostic amiante) et sa référence au sein du CSTB.
— Caractéristique diagnostic : est composée des informations extraites à partir des diagnostics. Elle contient le résultat de l'existence d'amiante dans un bâtiment similaire dans la même année, le code de ce bâtiment et l'année de sa construction.

Pour représenter les résultats des diagnostics effectués sur les produits, quand ils existent, nous avons modélisé le concept de Caractéristique diagnostic qui est décrit par le résultat de l'existence d'amiante propriété $A\_Diagnostic$ qui prend la valeur *oui* ou *non*, pour un code de bâtiment et une année de construction donnée. Les autres éléments de bâtiment (i.e. localisation, structure) et le bâtiment lui-même sont également décrits par une propriété $A\_Diagnostic\_Amiante$ qui prend la valeur *oui* ou *non* selon que l'un de ses éléments contient de l'amiante.

4 Approche prédictive

Dans les documents du CSTB, nous ne connaissons que les classes de produit utilisées pour une localisation mais nous ne connaissons pas la référence du produit utilisé. Par exemple, nous savons que le produit utilisé est un « Enduit » mais nous ne savons pas de quel enduit (i.e. produit commercialisé) il s'agit exactement.

Dans ce qui suit, nous proposons d'utiliser une méthode non-supervisée qui se base sur la présence de l'amiante dans les produits commercialisés pour prédire la probabilité de l'existence d'amiante pour les produits d'un bâtiment construit à une date donnée, puis plus globalement pour ses localisations, ses structures et le bâtiment lui-même. Dans une première étape, nous utilisons les ressources externes de l'ANDEVA (Association Nationale de Défense des Victimes de l'Amiante) et de l'INRS (Institut National de Recherche et de Sécurité) pour enrichir l'ontologie Amiante par des produits commercialisés, puis nous utilisons ces informations pour calculer la probabilité de présence d'amiante en utilisant un modèle de graphe probabiliste.

4.1 Enrichissement et peuplement de l'ontologie des produits commercialisés

Nous utilisons les deux ressources externes fournies par l'ANDEVA et l'INRS pour enrichir l'ontologie avec des sous-classes de produit et des instances de produits commercialisés avec leurs caractéristiques amiante. L'ANDEVA publie sur son site web [1] une liste des matériaux amiantés. Cette liste est représentée sous forme d'un tableau qui contient : la classe du produit, un nom ou un ensemble de noms de produits qui partagent les mêmes propriétés avec leur nom de fournisseur, et l'année à partir de laquelle ils ne sont plus amiantés. L'INRS publie également une liste [2] des matériaux amiantés représentés sous forme d'un tableau qui contient un ensemble de noms de produits, le nom éventuel du fournisseur, les intervalles de temps où le produit est amianté avec un degré d'incertitude, le type d'amiante, et les types d'utilisation. Plus précisément, dans la ressource INRS, les différents intervalles de temps peuvent correspondre à différentes annotations décrivant la présence d'amiante : le produit est amianté, la présence d'amiante est inconnue (non renseignée), le produit n'est plus amianté, le produit n'est plus commercialisé.

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

Dans un premier temps, nous extrayons automatiquement les informations concernant les produits de l'ANDEVA et l'INRS en nous basant sur la structuration des données et sur

1. http://andeva.fr/?-Liste-de-produits-contenant-de-l-
2. http://www.inrs.fr/media.html?refINRS=ED%201475

des expressions régulières pour l'extraction des intervalles et des probabilités de présence d'amiante dans les annotations. Puis, nous enrichissons l'ontologie avec les classes des produits qui deviennent des sous-classes de la classe $Produit$ de l'ontologie Amiante. Pour représenter les caractéristiques produits issues de chacune des sources, ainsi que les caractéristiques issues de la fusion de ces informations, nous utilisons la réification pour représenter les informations temporelles et les probabilités et ajoutons à l'ontologie Amiante représentée en figure 1 un nouveau concept de $Caracteristique\ Extraite$ qui va permettre de représenter les caractéristiques de présence d'amiante dans un matériau, en précisant l'intervalle de temps, la probabilité de la présence d'amiante, et la source de cette caractéristique (INRS, ANDEVA ou fusion) (cf. figure 2). Nous indiquons une probabilité de 1 pour les produits amiantés sur la période, de 0.5 quand la présence d'amiante est inconnu et de 0 sinon.

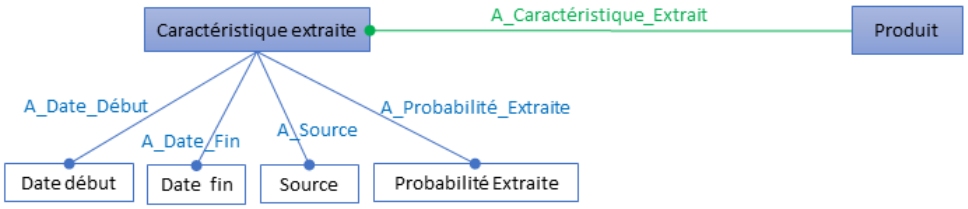

FIGURE 2 – *Représentation des caractéristiques extraites dans l'ontologie Amiante*

Prenons l'exemple de l'« ARMAZOL » qui est décrit dans l'INRS et l'ANDEVA par les tableaux 1 et 2 respectivement, et qui ne comporte pas de mention de fournisseur.

TABLE *1 – Extrait de l'INRS : « ARMAZOL »*

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

TABLE *2 – Extrait de l'ANDEVA : « ARMAZOL »*

Nom de famille de produits	Produit et fournisseur	Amianté jusqu'en
Revêtements de sols en dalles ou en rouleaux	ARMAZOL	1990

A l'issue de cette étape, la classe de produit « Revêtements de sols en dalles ou en rouleaux » est créée et deux instances de cette classe dont le nom est « ARMAZOL » sont créées qui ont les caractéristiques suivantes :

— Le premier produit est présenté comme provenant de la source INRS, et comme étant amianté (probabilité = 1) de 1946 à 1982, puis comme potentiellement amianté jusqu'en 1997 (probabilité = 0,5).
— Le second produit est présenté comme provenant de la source ANDEVA, il est amianté de 1946 à 1990 (probabilité = 1), puis non amianté (probabilité = 0) jusqu'en 1997.

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 (i.e. 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 se basant sur le nom commercial du produit (i.e. sur l'égalité de chaines de caractères qui ne diffèrent pas selon les deux sources).

Lors de la fusion des deux descriptions, nous fusionnons automatiquement les valeurs des propriétés d'un produit en faisant l'union des valeurs distinctes. Cependant, 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 faisons l'union ordonnée des bornes des intervalles de temps successifs de l'ANDEVA et l'INRS, en éliminant les doublons, et 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 calcul qui sera utilisé dans la prédiction de présence d'amiante dans le bâtiment.

Ainsi, la fusion des caractéristiques amiante du produit Armazol (schématisé en figure 3) conduit aux étapes suivantes :

1. Après avoir ordonné les bornes des intervalles d'entrée des deux sources, on obtient : {1946, 1982, 1990, 1997}.
2. A partir de ces bornes, nous construisons les intervalles du 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 ce conflit, nous prenons le maximum des deux valeurs (probabilité = 1 pour le premier intervalle et 0,5 pour le deuxième).

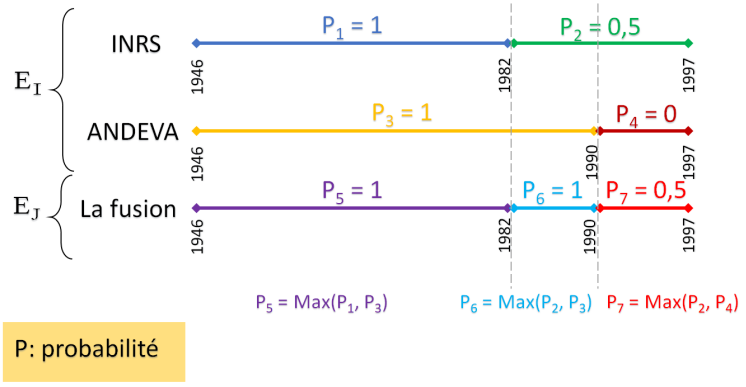

FIGURE 3 – *Exemple de fusion des intervalles de temps et des probabilités de présence d'amiante*

4.2 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.

TABLE *3 – Caractéristiques des produits appartenant à la classe "Revêtements muraux"*

Produit	Amianté jusqu'en
COLOVINYL	1983
DECOVER	/
NOVILON	1981
STRAIRLAM	/

TABLE *4 – Probabilités des produits de la classe "Revêtements muraux" en 1982*

Produit	Probabilité
COLOVINYL	1
DECOVER	0,5
NOVILON	0
STRAIRLAM	0,5

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

$$p_a(p_k, d) = \frac{\sum_{p_j \in F(p_k)} p_{ext}(p_j, d)}{\mid p_j \mid} \tag{1}$$

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 type F que p_k.

Par exemple : pour calculer la probabilité de l'existence d'amiante dans un produit de la classe de produits "Revêtements muraux" en 1982, nous allons utiliser les informations des ressources externes (ANDEVA et INRS) pour appliquer notre formule 1. La classe de produits "Revêtements muraux" contient les quatre produits décrits dans le tableau 3. Maintenant, nous remplaçons les variables dans l'équation 1, tel que $d = 1982$, $p_k = RevêtementsMuraux$ et le nombre de produits $\mid p_j \mid = 4$:

$$p_a(RevêtementsMuraux, 1982) = \frac{\sum_{p_j \in F(RevêtementsMuraux)} p_{ext}(p_j, 1982)}{4}$$

En 1983, nous disposons des probabilités de produits montrées dans le tableau 4 :
— "COLOVINYL" qu'est toujours amianté avant 1983, a la probabilité 1.
— Dans le cas de "DECOVER" et "STRAIRLAM" où nous n'avons pas des informations sur la présence d'amiante, et donc nous mettons la probabilité à 0,5.
— "NOVILON" n'est plus amianté à partir de 1981, donc sa probabilité est 0.
Nous trouvons alors :

$$p_a(RevêtementsMuraux, 1982) = \frac{1 + 0,5 + 0 + 0,5}{4} = 0,5$$

Cependant, l'une des difficultés est que l'INRS et l'ANDEVA se focalisent uniquement sur les produits ayant été amiantés 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. Nous proposons de le réajuster en se basant sur l'ensemble des diagnostics de prélèvement disponibles. Nous comparons pour cela, pour une année donnée d, la proportion de produits amiantés dans les ressources quelle que soit la classe de produit (i.e. $F(p_k) = Produit$) 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 par rapport à l'ensemble des diagnostics posés pour l'année. Cela nous permet de déterminer quelle est la proportion α de produits commercialisés manquants que nous considérons comme étant non amiantés.

$$p_a(p_k, d) \times (1 + \omega) = \frac{\sum_{p_j \in F(p_k)} p_{ext}(p_j, d)}{\mid p_j \mid + (\alpha \times \mid p_j \mid)} \tag{2}$$

Dans l'équation 2, 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 comme suit :

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

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 2, nous trouvons :

$$\alpha = \frac{\sum_{p_j \in F(p_k)} p_{ext}(p_j, d)}{\mid p_j \mid \times p_a(p_k, d) \times (1 + \omega)} - 1 \Rightarrow \alpha = \frac{-\omega}{1 + \omega} \tag{3}$$

Pour calculer la probabilité de présence d'amiante dans une localisation, une structure ou un bâtiment, nous considérons, comme l'expert, qu'il s'agit de la valeur maximum de l'ensemble des valeurs de probabilité de présence d'amiante des produits p_k qui composent la localisation l_i, puis de l'ensemble des localisations qui participent à une structure et enfin l'ensemble des structures composant le bâtiment. Ainsi, pour une localisation, nous appliquons :

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

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

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

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

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

Pour modéliser les probabilités amiante calculées (figure 4) dans l'ontologie Amiante, nous avons défini le concept de caractéristique calculée 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, l'année de calcul qui va correspondre à l'année du bâtiment et la classe de la probabilité. Seulement deux classes de probabilité ont été définies : forte, et faible (seuil à déterminer expérimentalement). En effet, compte tenu de l'incomplétude des données, une probabilité 1 ou 0 ne nous permet pas de prédire avec certitude que le bâtiment est ou n'est pas amianté.

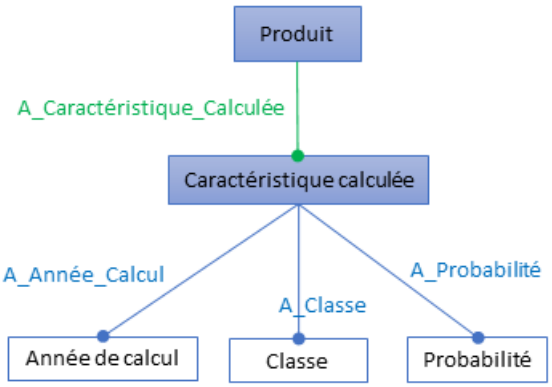

FIGURE *4 – L'ontologie Amiante avec les caractéristiques calculées*

5 Expérimentations

Nous avons mené une première expérimentation qui permet d'enrichir et de peupler l'ontologie Amiante en conservant les données origines et en calculant les caractéristiques des données fusionnées. Nous avons ensuite calculé la probabilité de présence d'amiante dans un bâtiment et de ces éléments pour un ensemble de données synthétiques que nous avons générées. L'objectif est d'observer l'évolution de cette probabilité au fur et à mesure des années sur les résultats obtenus.

5.1 Jeux de données

Description de l'ANDEVA et de l'INRS

Le fichier ANDEVA décrit 650 produits (28,2 Ko) 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 (28,1 Ko), 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 (i.e. classes de produit).

Génération de données synthétiques décrivant des bâtiments

Pour éviter d'obtenir des incohérences dans les données synthétiques décrivant les composants du bâtiment (ex. une localisation de type fenêtre ne peut contenir des produits de type ciment), nous avons défini un ensemble de contraintes que le processus de génération doit respecter. Ces 63 contraintes représentent des contraintes de domaine et de co-domaine pour les relations entre les structures et les localisations, et entre les localisations et les classes de produits.

La génération s'effectue de la manière suivante : nous créons un bâtiment dont la date de construction est aléatoirement choisie entre 1946 et 1997. Nous créons pour ce bâtiment un nombre aléatoire de structures que nous associons aléatoirement à une ou plusieurs localisations possibles en respectant les contraintes. De même pour chaque localisation générée nous

associons aléatoirement un ou plusieurs produits. Nous avons ainsi généré cent vingt (120) descriptions de bâtiments qui comportent 1133 produits.

5.2 Analyse quantitative des résultats

Enrichissement et Peuplement de l'ontologie avec l'INRS et l'ANDEVA

L'étape d'extraction de l'information et de fusion des produits commercialisés issus des deux ressources externes nous a permis de créer soixante-quatre (64) sous-classes de produit (e.g. enduit, colle, ...) et six cent quatre-vingt-quatorze (694) instances de produits commercialisés. Presque tous les produits de l'INRS sont également présents dans l'ANDEVA (256/300). Seulement 35 produits conduisent à des valeurs conflictuelles pour les caractéristiques amiantes.

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

Nous avons utilisé les diagnostics posés sur 40 produits pour calculer la valeur de α qui permet de réajuster le nombre de produits commercialisés. Les 40 produits diagnostiqués dont nous disposions concernent tous l'année 1963 pour laquelle $\omega \approx -50,19\%$ ce qui nous conduit après la résolution de l'équation 3 en utilisant les données disponibles, à $\alpha \approx 1,0076$.

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). Le graphe de la figure 5 montre l'évolution de cette probabilité pour l'ensemble des produits commercialisés. Ce graphe montre que la probabilité de présence d'amiante reste globalement stable jusqu'à 1972, puis elle décroit jusqu'à atteindre 0 en 1997, année où l'amiante est interdit. Sans le réajustement, la proportion de produits amiantés varie de $92,7\%$ à $44,8\%$. Quand le réajustement est appliqué, la proportion maximum de produits amiantés devient $46,17\%$. La figure 6 montre sur les quatre classes de produits présentées en tableau 5 que la probabilité de l'existence de l'amiante 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.

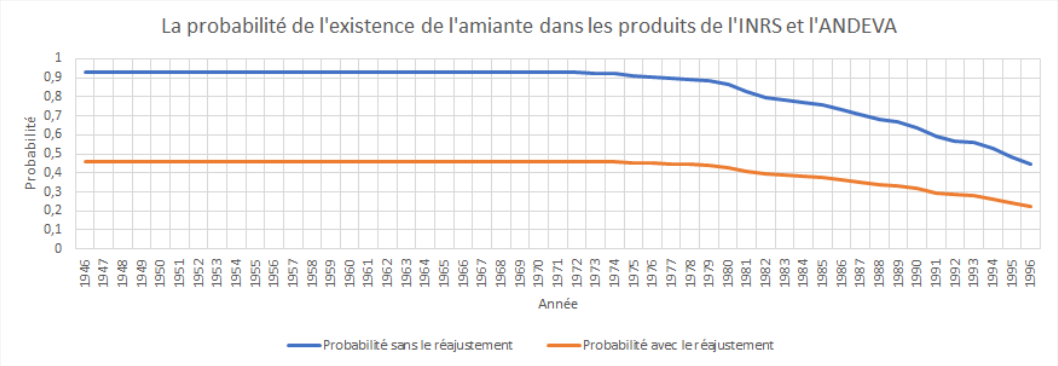

FIGURE 5 – *La probabilité de l'existence d'amiante dans les produits de l'INRS et l'ANDEVA en fonction des années*

La figure 7 montre la variation du nombre de bâtiments et leur probabilité moyenne de présence d'amiante en fonction des années dans les 120 descriptions de bâtiments générées. Comme les probabilités de présence d'amiante dans les produits sont propagées sur les éléments de bâtiments puis sur le bâtiment lui-même par une fonction maximum, les résultats ne montrent pas de baisse significative du nombre de bâtiments amiantés sur les données synthétiques. On peut ainsi remarquer que les 4 bâtiments datant de 1947 ont même probabilité moyenne que les 4 bâtiments datant de 1993 (probabilité de 0,2). Il suffit

TABLE 5 – *Caractéristiques des familles de produits choisis*

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

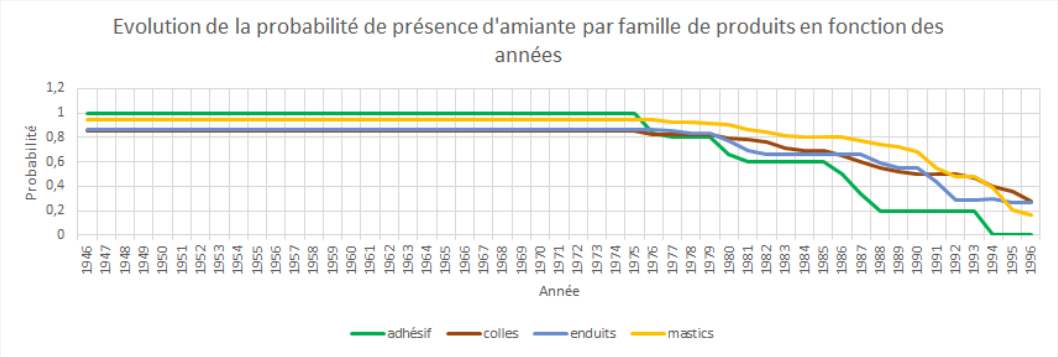

FIGURE 6 – *L'évolution de la probabilité de présence d'amiante par famille de produits en fonction des années*

en effet qu'une classe de produit utilisée soit peu désamiantée pour que le bâtiment entier soit considéré comme ayant un risque assez élevé de présence d'amiante.

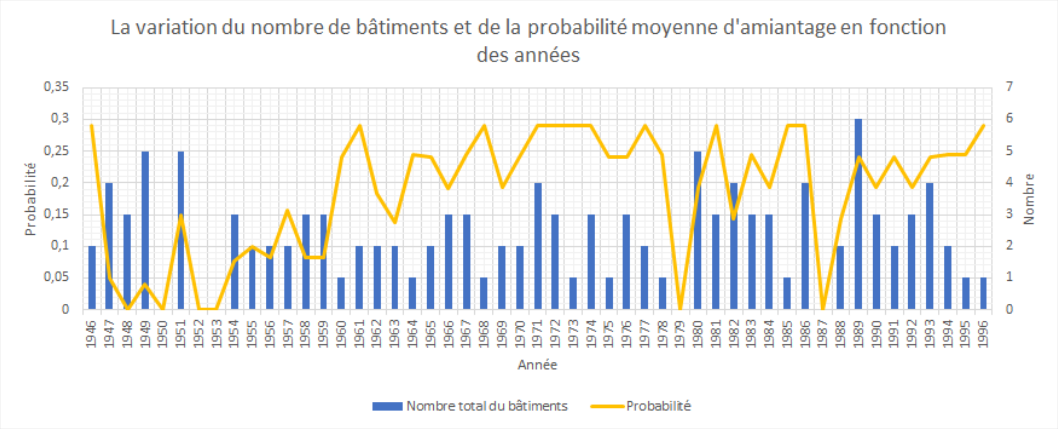

FIGURE 7 – *La variation du nombre de bâtiments et de la probabilité moyenne de présence d'amiante en fonction des années*

5.3 Analyse qualitative des résultats

Nous avons envoyé à l'expert un échantillon de résultats obtenus pour 30 bâtiments choisis aléatoirement pour qu'il valide le processus de calcul et les classes amiante de résultat. A l'origine, nos probabilités se répartissaient en quatre classes amiante (sans amiante, présence possible d'amiante, présence probable d'amiante, présence certaine de l'amiante). L'expert a validé le processus de calcul de probabilité à partir des produits commercialisés. Cependant, son expertise nous a conduit à ne considérer que deux classes amiante (faible et forte) car

l'incomplétude de données ne permet jamais de déduire une présence certaine ou une absence certaine d'amiante mais le seuil à utiliser reste à déterminer en utilisant d'autres diagnostics.

6 Travaux connexes

Approches de prédiction dans les graphes de connaissances

De récentes méthodes d'apprentissage supervisé cherchent à établir des prédictions en utilisant des données graphe (voir Hamilton *et al.* (2017a)) pour un état de l'art. L'objectif peut être de classer le rôle d'une protéine dans un graphe d'interaction biologique (Hamilton *et al.* (2017b)), de prédire le rôle d'une personne dans un réseau social, ou prévoir de nouvelles applications thérapeutiques de molécules de médicament existantes, molécules dont la structure peut être représentée sous forme de graphe. L'une des difficultés est alors de savoir représenter la structure du graphe, ou des éléments statistiques sur cette structure (e.g. degrés, coefficients de clustering) de façon à rendre celui-ci exploitable par des méthodes d'apprentissage. Aussi, certaines approches s'intéressent à l'apprentissage de ces modèles (Hamilton *et al.* (2017a)).
Les approches de programmation logique inductive et de découverte de règles (Lavrac (1994); Galárraga *et al.* (2015); d'Amato *et al.* (2016)) peuvent détecter des règles telles que *"Si une personne est née dans un pays, elle parle probablement la langue de ce pays"*. Ces règles peuvent ensuite être utilisées pour prédire de nouvelles informations et être syntaxiquement guidées par des motifs pour détecter des règles qui concluent sur une information ou une classe d'intérêt pour l'utilisateur (Nebot & Llavori (2012)). De telles règles apprises ont montré de bons résultats pour prédire des valeurs manquantes dans des graphes de connaissances généralistes (Galárraga *et al.* (2015)). Dans notre contexte, nous disposons de trop peu de données d'apprentissage pour apprendre à classifier automatiquement des éléments de bâtiments comme étant amiantés ou non amiantés, que ce soit par des méthodes d'apprentissage profond ou par des méthodes logiques basées sur la détection de règles. L'objectif ici est d'estimer une probabilité de présence d'amiante dans un produit inconnu appartenant à une classe de produit identifiée à partir des propriétés des produits commercialisés utilisés au moment de la construction du bâtiment. Cela revient à découvrir des règles qui, compte tenu de la classe d'un produit et de sa date, impliquent une présence d'amiante dans ce produit avec un certain degré de confiance. Les règles de propagation de cette probabilité à des éléments de bâtiments suivent ensuite une approche pessimiste (utilisation du maximum) qui reproduit le raisonnement d'un expert du bâtiment dans cette analyse de risque.

Incomplétude - Mesure de l'incomplétude

Comme les graphes sont généralement incomplets, les faits manquants ne doivent pas y être considérés comme faux (i.e. Hypothèse du monde ouvert - OWA). La notion de complétude d'un graphe de connaissance ne fait pas toujours sens selon les propriétés et les classes que l'on considère, et sans graphe de référence auquel se comparer (Razniewski *et al.* (2016)). Pour apprendre dans un tel contexte, certaines hypothèses complémentaires ont été introduites (Galárraga *et al.* (2016)). Dans (Galárraga *et al.* (2015)), une hypothèse de complétude partielle a été définie qui propose que, chaque fois qu'au moins un objet pour un sujet et une propriété donnés sont déclarés dans le graphe, tous les objets de cette paire sujet-propriété sont supposés être connus. D'autres approches ont pour objectif de mesurer l'incomplétude des données. Dans (Issa *et al.* (2017)), les auteurs proposent de calculer un schéma idéal à partir des propriétés fréquemment instanciées ensemble et d'en déduire les propriétés qui pourraient être considérées comme manquantes pour les instances de classe. Dans Tanon *et al.* (2018), les auteurs découvrent les cardinalités des propriétés dans les données. Dans notre approche, aucuns des liens entre les produits utilisés et les produits commercialisés ne sont disponibles, et il ne s'agit donc pas d'évaluer l'incomplétude des instances de cette propriété pour un bâtiment donné. Nous proposons d'évaluer le nombre de produits commercialisés non amiantés manquants, et donc le nombre d'instances de classe manquantes, en utilisant le petit nombre de diagnostics disponibles.

Fusion de données

Les approches de fusion de données ont proposé différents critères permettant de gérer les valeurs conflictuelles (Bleiholder & Naumann (2006); Mendes *et al.* (2012)). Certaines stratégies laissent l'utilisateur décider de la meilleure valeur à affecter à l'entité. D'autres stratégies sont automatiques et se basent sur des fonctions de filtrage qui choisissent la valeur la plus récente, la plus fréquente, la plus précise ou la plus fiable quand on dispose de la réputation des sources de données, ou sur des fonctions d'agrégation qui combinent les valeurs possibles en utilisant la moyenne, le minimum ou encore le maximum des valeurs possibles selon l'application. Dans notre approche, les conflits apparaissent lors de la fusion des probabilités de présence d'amiante dans les produits commercialisés associés aux intervalles de temps. Comme l'INRS et l'ANDEVA sont supposés avoir la même fiabilité et que nous ne disposons pas d'un grand nombre de sources qui permettraient d'envisager de se baser sur la fréquence d'une valeur, nous avons choisi de conserver les valeurs d'origine avec leur provenance, mais d'utiliser la fonction maximum dans la description fusionnée afin de tenir compte de ces deux sources pour évaluer de manière pessimiste le risque de présence d'amiante.

Graphes probabilistes

De nombreux travaux de recherche se sont intéressés à la modélisation, au requêtage et à la fouille de graphes probabilistes (Chekol *et al.* (2017); Benferhat *et al.* (2013)). Dans ces modèles, la probabilité peut être associée (1) aux arcs ou aux arêtes et représenter la probabilité d'existence d'un arc ou d'une arête entre deux nœuds du graphe,ou (2)aux nœuds et représenter la probabilité d'existence de ce nœud, ou (3) être associé aux attributs des nœuds ou des arcs (ou arêtes). Dans notre travail, nous ne représentons pas les liens d'identité qui relient potentiellement le produit inconnu utilisé dans un bâtiment à un produit commercialisé de la même classe, car nous considérons que chaque lien d'identité est équiprobable. Nous représentons de manière "réifiée" la probabilité de présence d'amiante dans les produits commercialisés et dans le produit inconnu (en utilisant le concept de caractéristique extraite ou de caractéristique calculée qui permet de lier un produit à une probabilité d'amiantage en conservant les données temporelles et la provenance de l'information pour les produits commercialisés).

7 Conclusion

Dans ce papier, nous avons présenté une première approche de prédiction de la présence d'amiante dans un bâtiment, approche qui devrait permettre d'aider un opérateur de repérage à décider des prélèvements qu'il est nécessaire d'effectuer dans un bâtiment construit à une date donnée. Nous avons tout d'abord défini une ontologie qui modélise les caractéristiques des bâtiments et les diagnostics quand ils existent. Cette ontologie est enrichie par des données temporelles probabilistes sur la présence 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.

Les premiers résultats montrent que cette probabilité évolue au fur et à mesure des années et qu'elle varie également en fonction des classes de produits utilisées dans un bâtiment.

Dans des travaux futurs, nous allons tester notre solution sur un ensemble de données réelles fourni par le CSTB. Nous planifions également d'utiliser cette probabilité calculée, la description des bâtiments, l'ontologie et un ensemble conséquent de diagnostics pour apprendre plus précisément les caractéristiques des bâtiments, des structures, des localisations et des produits qui peuvent influer sur la présence d'amiante dans les éléments de bâtiments, en utilisant la sémantique de l'ontologie.

Références

BENFERHAT S., KHELLAF F. & ZEDDIGHA I. (2013). A possibilistic graphical model for handling decision problems under uncertainty. In *8th conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-13)* : Atlantis Press.

BLEIHOLDER J. & NAUMANN F. (2006). *Conflict handling strategies in an integrated information system*. Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät ....

CHEKOL M. W., PIRRÒ G., SCHOENFISCH J. & STUCKENSCHMIDT H. (2017). Marrying uncertainty and time in knowledge graphs. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, p. 88–94.

D'AMATO C., TETTAMANZI A. G. B. & TRAN D. M. (2016). Evolutionary discovery of multi-relational association rules from ontological knowledge bases. In *Knowledge Engineering and Knowledge Management - 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings*, p. 113–128.

GALÁRRAGA L., RAZNIEWSKI S., AMARILLI A. & SUCHANEK F. M. (2016). Predicting completeness in knowledge bases. *CoRR*, **abs/1612.05786**.

GALÁRRAGA L., TEFLIOUDI C., HOSE K. & SUCHANEK F. M. (2015). Fast rule mining in ontological knowledge bases with AMIE+. *VLDB Journal*, **24**(6), 707–730.

HAMILTON W. L., YING R. & LESKOVEC J. (2017a). Representation learning on graphs : Methods and applications. *IEEE Data Eng. Bull.*, **40**(3), 52–74.

HAMILTON W. L., YING Z. & LESKOVEC J. (2017b). Inductive representation learning on large graphs. In *Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, p. 1025–1035.

ISSA S., PARIS P. & HAMDI F. (2017). Assessing the completeness evolution of dbpedia : A case study. In *Advances in Conceptual Modeling - ER 2017 Workshops AHA, MoBiD, MREBA, OntoCom, and QMMQ, Valencia, Spain, November 6-9, 2017, Proceedings*, p. 238–247.

LAVRAC N. (1994). Inductive logic programming. In *WLP*, p. 146–160 : Institut für Informatik der Universität Zürich.

MENDES P. N., MÜHLEISEN H. & BIZER C. (2012). Sieve : linked data quality assessment and fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, p. 116–123 : Citeseer.

NEBOT V. & LLAVORI R. B. (2012). Finding association rules in semantic web data. *Knowl.-Based Syst.*, **25**(1), 51–62.

RAZNIEWSKI S., SUCHANEK F. M. & NUTT W. (2016). But what do we actually know ? In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction, AKBC@NAACL-HLT 2016, San Diego, CA, USA, June 17, 2016*, p. 40–44.

TANON T. P., STEPANOVA D., RAZNIEWSKI S., MIRZA P. & WEIKUM G. (2018). Completeness-aware rule learning from knowledge graphs. In *IJCAI*, p. 5339–5343 : ijcai.org.