D'un modèle statistique à un modèle de connaissance : retour d'expérience

Rabia Azzi, Sylvie Despres, Jérôme Nobécourt

D'un modèle statistique à un modèle de connaissance : retour d'expérience

Rabia Azzi[1], Sylvie Despres[1], Jérôme Nobecourt[1]

UNIVERSITÉ PARIS 13, SORBONNE PARIS CITÉ, LIMICS,(U1142), INSERM, Sorbonne Universités, UPMC
Université Paris 6, 74 rue Marcel Cachin F-93017 Bobigny cedex, France
prenom.nom@univ-paris13.fr

Résumé :
Les modèles statistiques sont couramment représentés sous forme textuelle, tabulaire et graphique dans des documents (rapports, articles, affiches et présentations) qui sont le plus souvent en format PDF. Même si ce format rend l'accès à l'information plus difficile, il est intéressant de traiter directement le fichier PDF. Dans cet article nous proposons une approche permettant le passage d'un modèle statistique de connaissances à un modèle de connaissances qui soit visualisable afin d'en permettre une exploitation plus aisée. Notre approche consiste à : (i) extraire les informations pertinentes sous forme de triplets RDF ; (ii) organiser les triplets pour construire un modèle conceptuel ; (iii) visualiser dynamiquement le modèle obtenu. Nous nous focalisons sur les deux premières étapes de la méthodologie.

Mots-clés : Modèle statistique, Tableau statistique, Modèle conceptuel, Extraction sémantique d'information, RDF

1 Introduction

Dans cet article, nous traitons de l'automatisation de l'interprétation de résultats d'approches statistiques publiées au format PDF. Les questions soulevées sont : (1) comment extraire l'information et la sémantique existante dans ces documents ; (2) comment traduire les informations extraites vers un modèle conceptuel ; (3) comment exploiter le modèle obtenu pour le rendre dynamique et interactif.

La démarche statistique, indispensable pour extraire des connaissances à partir de données, est utilisée dans presque tous les domaines de l'activité humaine : ingénierie, management, économie, biologie, informatique, etc. (Saporta, 2011). Les données sont collectées, exploitées, analysées et enfin souvent présentées sous formes tabulaires et graphiques afin de faciliter leur interprétation.

Un tableau statistique est un ensemble de cellules organisées en lignes et colonnes, contenant des données chiffrées. Un tel tableau peut être : (1) à une entrée, il permet l'étude d'un caractère d'une population ; (2) à double entrée, il permet d'étudier simultanément deux caractères d'une population. Quelle que soit la structure du tableau, l'approche de lecture des informations qu'il contient est identique. Il faut dans tous les cas étudier le titre, la source et comprendre les intitulés des lignes et des colonnes. Les informations représentées dans les tableaux sont fortement connectées, en particulier les liens entre les éléments du tableau sont implicites et nécessitent une explicitation. Cependant, comme toutes les données sont d'égale importance, il n'est pas évident d'identifier et de sélectionner les informations essentielles (Junyong & Sangseok, 2017). Notre objectif est de proposer un traitement semi-automatique pour l'interprétation de ces tableaux. En effet, ils constituent un des moyens les plus couramment utilisés pour présenter et structurer les informations. Ce traitement nécessite d'extraire les données qu'il contiennent et de les représenter dans un modèle facilitant leur compréhension. En outre, la plupart des documents sont publiés au format PDF à partir duquel l'extraction d'information est fastidieuse. De nombreuses approches ont été proposées pour extraire des informations à partir de documents PDF. Elles exploitent, soit le format des balises tel que HTML et XML, soit un format de texte brut pour l'extraction d'information (Riaz *et al.*, 2016), (Klampfl & Kern, 2015).

Dans cet article, nous décrivons une méthode permettant de traduire des tableaux statistiques au format PDF vers un modèle conceptuel guidant l'interprétation des informations

qu'il contient. Nous utilisons les techniques d'extraction d'information à partir de documents PDF et le modèle de triplets RDF [1] afin de structurer les informations extraites. Cette approche permet en effet d'explorer la manière de représenter des informations sans identifier initialement un vocabulaire source pour les prédicats (Powell, 2015). D'un point de vue applicatif, RDF utilise des identifiants uniques pour les ressources et chaque triplet correspond à la déclaration d'un fait.

Cet article est organisé comme suit : dans la section 2, nous introduisons la démarche de présentation des résultats en statistique. Dans la section 3, nous présentons l'approche générale et nous nous focalisons sur les deux premières étapes de la méthodologie (extraction des connaissances sous forme de triplets RDF et construction d'un modèle conceptuel). Dans la section 4 et 5, nous décrivons l'expérimentation, les résultats obtenus et leur évaluation. Dans la section 6, nous concluons et présentons des perspectives d'investigations complémentaires.

2 Présentation des résultats en statistique

Le mot statistique [2] désigne à la fois un ensemble de données d'observations et l'activité consistant en leur recueil, leur traitement et leur interprétation. Pour (Saporta, 2011), faire de la statistique suppose l'étude d'un objet ou d'un ensemble d'objets sur lesquels des caractéristiques appelées « variables » sont observées. La notion fondamentale en statistique est celle de population qui correspond à un groupe ou à un ensemble d'objets équivalents. Les objets sont appelés individu ou unité statistique. L'étude de tous les individus d'une population finie correspond à un recensement. Lorsqu'une partie de la population est observée, il s'agit d'un sondage, la partie étudiée étant désignée comme l'échantillon. Ainsi, chaque individu d'une population est décrit par un ensemble de caractéristiques appelées variables ou caractères. Ces variables peuvent être classées comme des variables quantitatives ou numériques (par exemple, taille, poids, etc.) ou comme des variables qualitatives s'exprimant par l'appartenance à une catégorie (par exemple, catégorie socio-professionnelle, etc.).

La démarche statistique comporte usuellement les étapes suivantes :

— le recueil qui consiste à collecter les données, selon deux grandes méthodologies : les sondages et les plans d'expériences ;
— l'exploration qui consiste à synthétiser, résumer, structurer l'information contenue dans les données ;
— l'inférence qui consiste à étendre les propriétés constatées sur l'échantillon de la population toute entière et à valider ou infirmer des hypothèses *a priori* ou formulées après une phase exploratoire ;
— la modélisation qui consiste généralement à rechercher une relation approximative entre une variable et plusieurs autres. Les modèles souvent utilisés sont la régression linéaire, le modèle linéaire général et la méthode de discrimination.

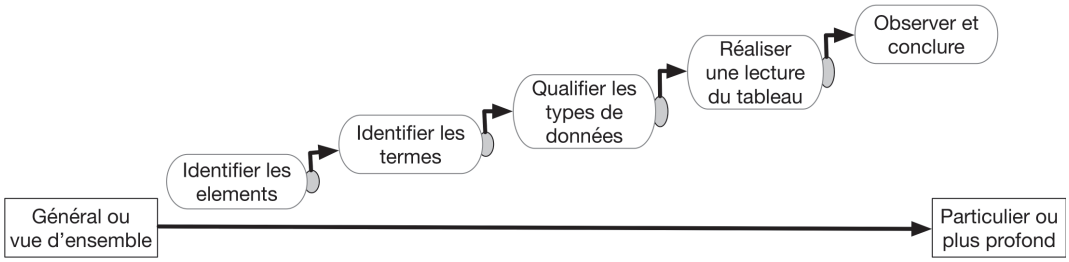

FIGURE *1 – Etapes d'analyse d'un tableau statistique*

1. https ://www.w3.org/RDF/
2. http ://www.larousse.fr/dictionnaires/francais/statistique/74516

Les méthodes visuelles sont largement utilisées en statistique pour présenter les résultats de manière claire et concise. Il existe plusieurs formes possibles de présentation visuelle des résultats : tableaux, graphiques, histogrammes, diagrammes, etc. Parmi ces formats, les tableaux sont les plus utilisés, car l'information est disposée de manière à mettre en évidence les relations entre les données.

Il existe trois types de tableaux : (i) les tableaux de données (les premiers construits) qui sont généralement « de grande taille » puisqu'ils comptent autant de lignes que de sujets étudiés ; (ii) les tableaux de distribution de variables (les plus connus) sont obtenus par regroupement des cases identiques figurant dans les colonnes et décrivent la distribution d'une variable ; (iii) les tableaux de contingence sont constitués par croisement de deux variables renseignées.

Les tableaux statistiques permettent d'organiser et de présenter les données ou les résultats en regroupant des informations de même nature. Cependant pour les exploiter, une démarche rigoureuse doit être suivie. Le grand principe d'analyse d'un tableau en statistique (voir Figure 1) est d'adopter une démarche allant du général au particulier. Cette démarche comporte les étapes suivantes :

— identifier les éléments : consiste à identifier le titre, la source de l'étude, la nature du tableau, etc. Chacun de ces éléments est porteur d'information (par exemple, le titre peut renseigner sur l'idée, la variable expliquée, etc.) ;
— identifier les termes : consiste à identifier les termes figurant dans le titre, les colonnes et les lignes du tableau, etc. ;
— qualifier les types de données : consiste à qualifier le type de données contenu dans le tableau en prenant en compte les unités (par exemple, des pourcentages, des probabilités, etc.) ;
— réaliser une lecture du tableau : consiste à appliquer deux règles communes de lecture à tous les tableaux. La première règle consiste à construire une paraphrase en débutant la lecture en se plaçant sur une ligne et en poursuivant par celle des colonnes utiles à l'analyse. La seconde consiste à répéter la première règle sur plusieurs lignes du tableau pour vérifier la pertinence des relations ;
— observer : consiste à tirer des conclusions à partir du tableau. Par exemple, identifier des relations entre variables (cause/effet), des valeurs extrêmes, des tendances, etc.

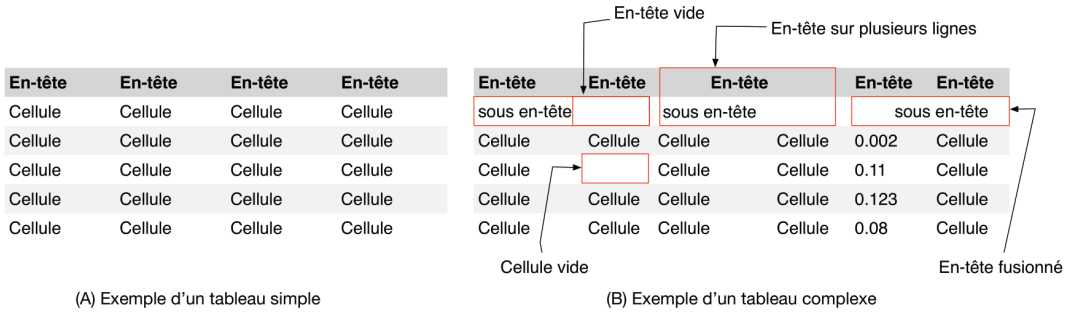

FIGURE *2 – Exemple de tableau simple et complexe*

Si une analyse visuelle permet à un être humain de reconnaître et comprendre facilement les tableaux, la situation est différente quant il s'agit d'un ordinateur. Un tableau est constitué de cellules d'en-tête et de cellules contenant des données. Cette structuration permet de définir la relation entre les cellules et fournit un contexte aux utilisateurs. Selon (Yeon-Seok & Kyong-Ho, 2008), il existe deux types de tableaux (voir Figure 2) : (1) les tableaux simples (voir Figure 2-A) comportant au maximum une ligne d'en-tête subdivisée en colonnes d'en-têtes. La colonne d'en-tête précise le type d'information qu'elle contient. Il n'y a pas de cellules fusionnées dans un tableau simple ; (2) les tableaux complexes (voir Figure 2-B), sont constitués d'en-têtes composées d'une ou plusieurs lignes ou de plusieurs cellules. Plusieurs lignes peuvent être associées à une même cellule d'en-tête. Elles peuvent également

contenir des cellules vides et des cellules fusionnées.

Prenons comme exemple, le tableau (B) de la Figure 2, l'analyse visuelle de ce tableau permet d'identifier les difficultés suivantes :

— les en-têtes des colonnes correspondent généralement aux variables utilisées dans l'étude (par exemple, le nombre de cas, la probabilité, etc) mais elles peuvent figurer sur plusieurs lignes et comportent parfois des cellules vides ;
— les lignes sont composées de cellules qui peuvent parfois être vides. Généralement, le contenu de la première cellule correspond à une variable de l'étude statistique et le reste à la valeur de l'association entre l'en-tête des colonnes et de la ligne ;
— les contenus des cellules peuvent être différents (texte, chiffres, caractères spéciaux, etc.).

L'approche proposée dans cet article prend en compte à la fois la démarche d'analyse d'un tableau statistique en tenant compte des difficultés précédemment identifiées.

3 Approche proposée

Nous avons conçu une application qui repose sur l'approche décrite en Figure 3. Le traitement prend en entrée un modèle statistique au format PDF, et se décompose en 3 étapes :

E1 : extraction de connaissances sous forme de triplets RDF.

E2 : construction d'un modèle conceptuel.

E3 : visualisation dynamique du modèle conceptuel.

Dans cet article, nous présentons l'approche permettant la traduction du modèle statistique vers le modèle conceptuel (E1 et E2).

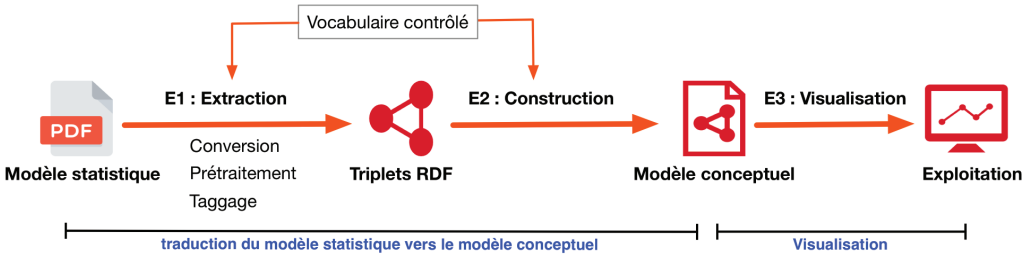

FIGURE *3 – Etapes de la méthodologie générale*

3.1 Extraction des connaissances (E1)

Cette section présente un état de l'art relatif à l'extraction de connaissances (3.1.1) ainsi que l'approche d'extraction de connaissances sous forme de triplets RDF à partir du document PDF (3.1.2 et 3.1.3).

3.1.1 Etat de l'art

Ces dernières années, le web est devenu un référentiel universel de données. L'accessibilité de ces données a permis l'émergence de nombreuses approches d'extraction de connaissances à partir de données structurées et non structurées (Unbehauen *et al.*, 2012). Souvent, le flux de données traitées par ces approches d'extraction est issu de tables au sens HTML. L'extraction de ces connaissances permet d'alimenter un grand nombre d'applications (Crestan & Pantel, 2010). La richesse et l'utilité des tables relationnelles sur le web ont permis l'émergence de plusieurs approches d'extraction de connaissances sous forme de triplets RDF (Muñoz *et al.*, 2014), (Lu *et al.*, 2015) ou sous forme d'une description logique (Pivk *et al.*, 2007).

D'autres format sont également utilisés pour coder des tables, le format PDF est le plus répandu. Pour extraire et exploiter ces contenus, il est nécessaire de mettre en place des approches appropriées (Ronzano & Saggion, 2016). Les approches mises en place sont personnalisées et sont fondées sur des éléments structurels tels que le titre, les sections, les figures, les tableaux, etc. (Riaz *et al.*, 2016), (Wu *et al.*, 2015). Cependant, pour exploiter les documents au format PDF, il est nécessaire de les convertir vers un format exploitable (XML, Textuel, HTML, etc.).

Dans cet article, nous nous concentrons sur l'extraction de triplets RDF à partir d'études statistiques publiées au format PDF. Le document PDF est converti dans un format exploitable. Puis, à l'aide d'un vocabulaire contrôlé, les triplets RDF sont extraits à partir des tables HTML.

3.1.2 Conversion du fichier PDF

Le format PDF est devenu une norme du support de lecture numérique (ordinateurs, liseuses, tablettes, smartphones, PDA, etc.). L'objectif initial du PDF était de préserver et protéger le contenu et la mise en page d'un document, quels que soient la plate forme ou le programme informatique dans lequel il est visualisé. C'est pourquoi, les fichiers PDF sont difficiles à modifier et parfois même, l'extraction d'information à partir de ces fichiers constitue un véritable défi.

	Plateforme	**Licence**	**Langage de programmation**	**Mise en page conservée**	**Format de sortie**	**Dernière mise à jour**	**Extraction de table**	**Traitement**
Tabula	Application et service web	MIT	Java	Sélection manuelle	JSON CSV	27/02/2018	Oui *via* l'intervention de l'utilisateur	Semi-automatique
PDFMiner	Ligne de commande	MIT	Python	Non	Text, HTML	24/03/2014	Non	automatique
Pdftohtml	Ligne de commande	Copyright	C	Non	HTML, XML	03/08/2006	Non	automatique
Pdf2htmlEX	Ligne de commande	GPLv3	Python	Oui	HTML	11/12/2016	Non	automatique
PDFX	Service web et ligne de commande	-	JavaScript	Non	HTML,XML	-	Non	automatique
PDF Online	Service web et Api	Copyright	-	Oui	HTML	-	Non	automatique

FIGURE *4 – Tableau comparatif des outils de conversion de documents PDF*

En outre, la forme des fichiers PDF varie, ce qui conduit à la mise en place de méthodes de traitements adaptées à chacune d'entre elles. Dans tous les cas, pour automatiser l'extraction d'information, il convient de convertir ces fichiers dans un format exploitable par la machine. Plusieurs outils ont été développés pour aider ce processus de conversion. Pour justifier le choix de l'outil que nous avons utilisé, nous proposons une analyse de ceux souvent cités comme référence :

— pdftohtmlEX [3] convertit les fichiers PDF au format HTML en conservant le texte et la mise en forme des tableaux ;
— pdftohtml [4] convertit les fichiers PDF au format HTML et XML ;
— PDFX [5] utilise des règles pour reconstruire la structure logique d'articles scientifiques au format PDF, quels que soient leur style de formatage ;
— Tabula [6] extrait semi-automatiquement des tableaux de données à partir de fichiers PDF ;
— PDFMiner [7] extrait des informations à partir de documents PDF. Contrairement à d'autres

3. http ://coolwanglu.github.io/pdf2htmlEX/
4. http ://pdftohtml.sourceforge.net/
5. http ://pdfx.cs.man.ac.uk/
6. http ://tabula.technology/
7. http ://www.unixuser.org/ euske/python/pdfminer/

outils liés au PDF, il est entièrement consacré à l'obtention et l'analyse de données de texte ;
— PDF Online [8] convertit les fichiers PDF au format HTML en conservant le texte et la mise en forme des tableaux.

Une synthèse des caractéristiques de ces outils est présentée dans le tableau de la Figure 4. Nous avons sélectionné l'outil pdftohtmlEX (voir Figure 5), pour les raisons suivantes : (1) le format de sortie HTML préserve la structure tabulaire lors du processus de conversion ; (2) l'information concernant la position du tableau dans le document est présente. Ainsi, il est possible de refaire l'action inverse, par exemple en partant des éléments extraits il est possible d'identifier l'endroit du tableau contenant cette information ; (3) la conversion est complètement automatique contrairement à l'outil Tablula. En revanche, la sortie HTML obtenue présente l'inconvénient d'être linéaire et nécessite des traitements supplémentaires.

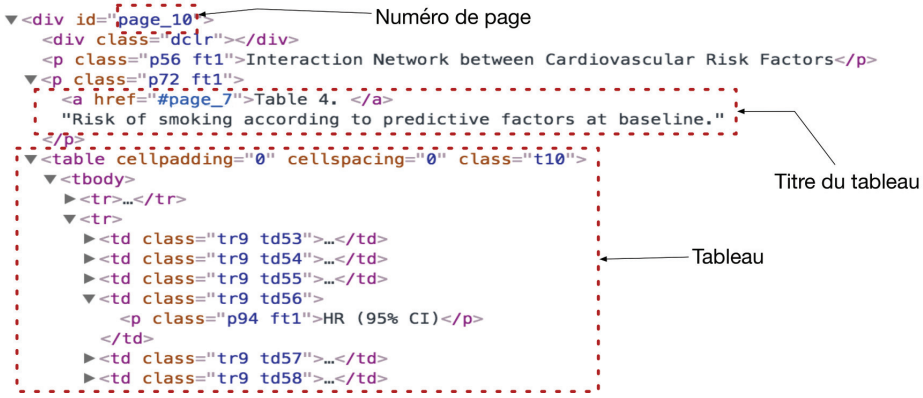

FIGURE 5 – *Sortie HTML obtenue avec pdftohtmlEX*

3.1.3 Localisation et extraction de l'information

3.1.3.1 Localisation de l'information

Une des difficultés de ce travail concerne la reconnaissance de tableaux contenant des informations relatives à un contexte d'étude. Il est facile pour un humain d'identifier les tableaux pertinents dans ce contexte. Par exemple pour un humain, l'analyse visuelle d'un tableau peut lui permettre de déduire facilement le sujet étudié. Ce processus de déduction est réalisé en identifiant certains termes dans le titre, légende, etc. L'automatisation de ce processus est souvent non triviale, même lorsque le système peut localiser les tableaux par reconnaissance de balises « TABLE » dans le document HTML. Le problème à résoudre est : comment déterminer automatiquement si les informations décrites dans un tableau concernent le contexte de l'étude ?

Plutôt que de localiser les informations de manière similaire à (Ermilov *et al.*, 2013) ou de trouver des heuristiques comme l'ont suggéré (Shigarov, 2015) et (Clark & Divvala, 2015), nous avons opté pour une approche supervisée exploitant un vocabulaire contrôlé s'il existe ou à partir des données de l'étude. Nous identifions les chaînes de caractères contenant les termes du vocabulaire. Nous pouvons, par exemple, déterminer qu'un tableau est pertinent s'il contient les termes du vocabulaire dans le titre et/ou le corps du tableau. Contrairement aux deux approches qui supposent un travail important d'élaboration d'heuristiques et des règles, notre démarche est générique et ne nécessite qu'un pré-traitement léger.

8. http ://www.pdfonline.com/

3.1.3.2 Extraction de l'information

Après avoir identifié les tableaux pertinents (contenant des informations sur le contexte de l'étude) dans le document HTML, l'étape suivante consiste à extraire ces informations et les liens qui leur sont associés.

L'extraction des informations à partir d'un tableau nécessite deux étapes : (1) la détermination des colonnes pertinentes qui est obtenue en utilisant un vocabulaire contrôlé ; (2) la construction des paraphrases d'interprétation des données qui est obtenue en prenant en compte le titre de chaque ligne, les en-têtes des colonnes pertinentes et les cellules se trouvant à l'intersection entre la ligne et l'en-tête.

Notre approche d'extraction comprend quatre étapes :

1. **L'extraction des tableaux pertinents** comprend deux phases : (a) la reconnaissance des tableaux ; (b) la vérification de la pertinence du tableau. La reconnaissance des tableaux est utilisée pour identifier et extraire tous les tableaux présents dans le document en utilisant le format des balises HTML. La vérification de la pertinence des tableaux a pour objectif de déterminer si le tableau extrait traite du sujet d'étude. Le principe consiste à taguer chaque terme du vocabulaire reconnu dans le tableau. Ainsi, s'il contient des termes du vocabulaire, il est déclaré pertinent et stocké sous forme d'un tableau associatif ($numero\_de\_page => titre => contenu$). Dans le cas contraire, il est rejeté.
2. **L'extraction des colonnes pertinentes** consiste à extraire les colonnes dont les en-têtes contiennent les termes du vocabulaire. Cependant, les en-têtes des tableaux peuvent apparaître sur plusieurs lignes avec des en-têtes vides ce qui rend la tâche plus difficile. Pour résoudre ce problème, nous avons commencé par identifier les modèles d'en-têtes utilisés dans les tableaux. Puis, nous avons construit des règles d'extraction pour ces modèles. Par exemple, si l'en-tête de la colonne comporte deux lignes alors la règle de traitement consiste à dupliquer le titre de la première ligne dans la seconde ligne. Souvent les tableaux extraits comportent des cellules vides et des lignes non pertinentes. Un traitement de nettoyage et une mise en forme des tableaux doit par conséquent être appliqué.
3. **Le nettoyage et la mise en forme des tableaux** se déroulent en trois étapes :
   — **substitution des cellules vides** : les cellules vides sont dues à la structuration des tableaux sur plusieurs niveaux. Le traitement permet de recopier le contenu de la cellule précédente dans les cellules vides suivantes d'une même colonne. Pour réaliser ce traitement, on doit considérer le type des données des cellules afin de ne pas introduire de biais. Par conséquent, ce traitement concerne uniquement les cellules contenant des chaînes de caractères.
   — **suppression de lignes non pertinentes** : les lignes non pertinentes proviennent de la structuration des tableaux qui fournissent des informations supplémentaires (telles que, l'année de l'étude, la source, etc.). Le traitement permet de vérifier deux paramètres pour chaque ligne du tableau, la présence des termes du vocabulaire et la présence de plusieurs cellules vides. En combinant ces deux paramètres, les lignes non pertinentes sont automatiquement supprimées.
   — **renommage d'en-tête** : certaines cellules d'en-têtes peuvent être vides ce qui constitue un obstacle à la lecture de l'information. Nous avons construit des règles à l'issue de l'analyse de la structure des tableaux. Par exemple, si le titre de la ligne i du tableau T tient sur deux cellules (C_1 et C_2) alors la valeur des en-têtes pour C_1 et C_2 sera respectivement « Class » et « Label ». Le traitement appliqué remplace respectivement l'en-tête vide par les valeurs « Class » et « Label ».

   À l'issue de ces trois étapes, on obtient une collection de tableaux nettoyés et exploitables.
4. **La construction de triplets** : résulte de l'extraction des informations réalisée sur la collection des tableaux T. Le résultat obtenu est un ensemble de triplets appelé graphe RDF décrivant l'ensemble des tableaux pertinents.

L'approche choisie est décrite dans la Figure 6 (A). L'automatisation du processus de construction des paraphrases est réalisée de la manière suivante :
Un tableau T est décrit par un ensemble de triplets RDF représentant les lignes le constituant. Chaque ligne i du tableau est considérée comme un nœud blanc noté _ :x_i. Une ligne i est décrite par un ensemble de triplets $(s_i, p_k, o_{i,k})$, où :
— s_i : correspond au nœud blanc _ :x_i associé a la ligne i ;
— p_k : correspond à un littéral, à partir du contenu de l'en-tête de la colonne k ;
— $o_{i,k}$: correspond à un littéral typé, à partir du contenu de la cellule à l'intersection entre la ligne i et la colonne k.
Chaque ligne i d'un tableau T comporte un titre. Ce titre peut être présent : (1) sur une colonne, dans ce cas, la première colonne sera décrite par un seul triplet ; (2) sur deux colonnes fusionnées, dans ce cas, la première colonne sera décrite par deux triplets. Les en-têtes des colonnes les plus répandus dans les tableaux statistiques sont la probabilité et le nombre d'individus pour chaque variable étudiée. En partant de ce constat, nous avons construit un modèle générique de triplet guidant leur extraction (voir Figure 6 (B)) . Dans ce modèle, chaque ligne du tableau va comporter un « Label », une « Class », un « Nombre de cas » et une « Probabilité ». Les littéraux typés sont dans un premier temps considérés comme des chaînes de caractères mais seront par la suite décomposés en un « réel » et une « unité » afin éviter une perte d'information sur la donnée.
Puis, chaque titre F d'un tableau T est segmenté pour identifier les éléments décrits et le terme désignant la relation permettant de relier les éléments du titre F au tableau T. Ce traitement est réalisé à l'aide de TreeTagger [9]. A partir de l'élément et de la relation identifiées dans le titre F, des triplets sont construits selon le modèle (s, p, o) où :
— s correspond a l'élément identifié dans le titre ;
— p correspond au type de relation entre le titre et le tableau ;
— o correspond au nœud blanc « _ :x_i » pour chaque ligne du tableau T.

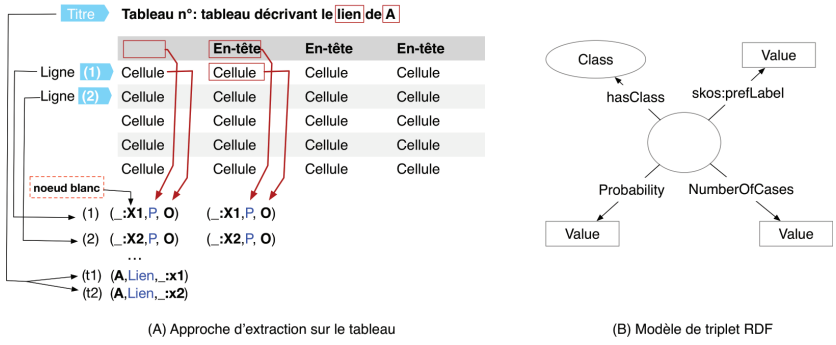

FIGURE 6 – *Stratégie d'extraction des triplets à partir des tableaux*

3.2 Construction du modèle conceptuel (E2)

Une des limites des modèles statistiques est qu'ils n'apportent qu'une information brute et faiblement structurée par rapport au contexte d'étude. En effet, la description et l'interprétation d'une relation dans un tableau sont construites sur des observations soumises à la subjectivité de l'observateur. Par exemple, deux personnes ayant des niveaux de connaissances différents sur un sujet n'auront pas la même interprétation d'un tableau.

L'approche proposée pour réduire ce biais consiste à construire un modèle conceptuel du domaine étudié pour guider l'interprétation des connaissances implicites dans les tableaux. L'avantage de ce modèle est : (i) de permettre d'avoir une description structurée ; (ii) d'être souple en permettant d'étendre à d'autres concepts du domaine.

9. http ://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

4 Expérimentation

Ce travail s'inscrit dans le contexte de l'amélioration de la prévention du risque cardiovasculaire qui cause près de 17,7 millions de décès, soit 40% de la mortalité mondiale totale (WHO, 2017). Si les principaux facteurs de risques cardiovasculaires sont aujourd'hui bien connus, leur évaluation à tendance à être réalisée sans considérer les interactions qui les lient (Meneton *et al.*, 2016).

Pour conduire notre expérimentation, nous travaillons à partir d'une étude statistique menée par (Meneton *et al.*, 2016) dans le domaine de l'épidémiologie. Son objectif était de mettre en évidence des interactions entre des facteurs de risques cardiovasculaires. Les résultats obtenus par les auteurs en appliquant des tests de régression sont résumés dans des tableaux, figures et textes.

Le but de notre expérimentation est double : extraire les connaissances contenues dans les tableaux sous forme de triplets RDF et construire le modèle conceptuel à partir de ces triplets.

L'étude de (Meneton *et al.*, 2016) est publiée sous forme d'un document PDF dans lequel 13 tableaux décrivent les associations de 13 facteurs de risque cardiovasculaires. La première étape était de convertir le fichier PDF vers un format exploitable à l'aide de pdftohtmlEX. Le résultat obtenu est un document HTML contenant « 9798 lignes », « 55702 mots », « 1254107 caractères » et « 32 tableaux ».

Pour illustrer notre démarche, nous avons choisi un tableau (voir Figure 7) du document PDF. L'ensemble des tableaux décrivant les associations entre les facteurs de risque cardiovasculaires ont tous la même structure.

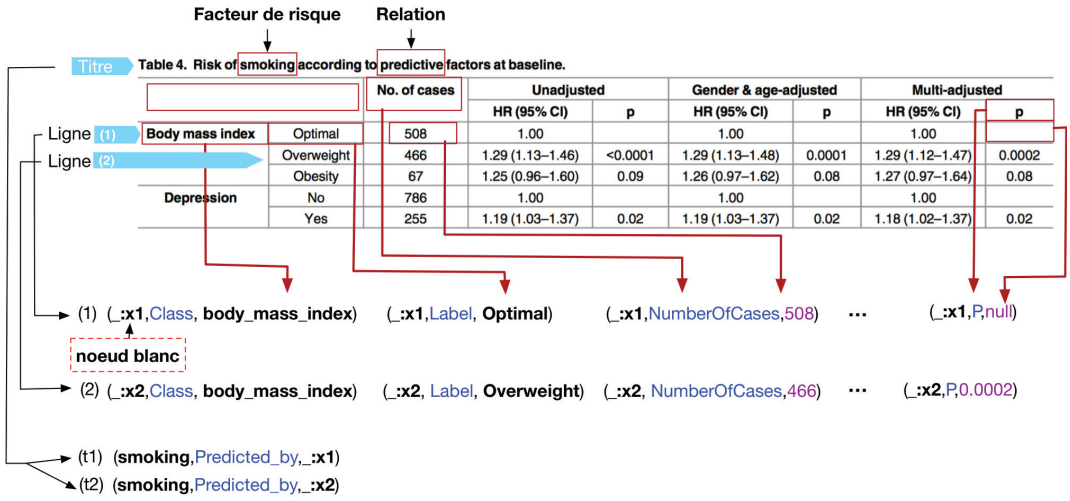

FIGURE 7 – *Stratégie d'extraction de l'information appliquées sur le tableau*

Chaque ligne i du tableau (Figure 7) comporte un titre. Ce titre est écrit sur deux colonnes fusionnées. Le titre de la première ligne du tableau est composé de « Body mass index » et « Optimal ». Dans ce cas, la première colonne sera décrite par deux triplets de la forme $(\_ :\mathrm{x_i}, p, o)$, où :

— « _ :$\mathrm{x_i}$ » correspond à la valeur du nœud blanc ;
— « p » correspond respectivement à « Class » et « Label » pour la première colonne ;
— « o » correspond à la valeur de l'intersection entre la ligne i et la colonne k du tableau T. Les contenus des colonnes « P » (décrivant la probabilité) et « No. of cases » (décrivant le nombre d'individus pour chaque Class) sont traités comme des chaînes de caractères.

Chaque ligne du tableau est décrite par un triplet suivant le modèle décrit dans la Figure 8-A . Par exemple, le résultat obtenu pour la troisième ligne du tableau (A) est présenté Figure 8-B.

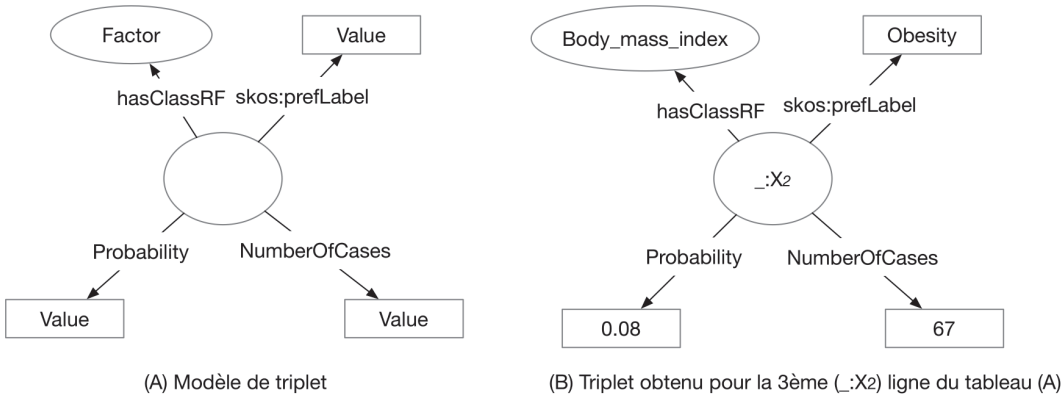

FIGURE *8 – Modèle de triplet pour l'extraction d'information*

— **Construction du modèle conceptuel (E2)** :
«CARRE [10] » est à notre connaissance le seul vocabulaire actuellement disponible. Il décrit les facteurs de risque cliniques cependant, il contient peu de termes relatifs aux facteurs de risque cardiovasculaires et est difficilement réutilisable. Pour pallier cette incomplétude, nous avons élaboré un vocabulaire contrôlé en utilisant les termes du domaine des maladies cardio-vasculaires utilisés par les experts du domaine. Le modèle conceptuel obtenu (voir Figure 9-(2)) a été élaboré en utilisant ce vocabulaire contrôlé. Dans ce modèle, chaque facteur de risque est décrit par sa catégorie, les facteurs qui le prédisent, le label préféré et le label alternatif. Par exemple, le facteur de risque « Smoking » est décrit par :

1. catégorie des « Behavioral_factors » ;
2. prédit par les facteurs associés aux nœuds blancs ($\_ :x_1, \_ :x_2, \_ :x_3$) ;
3. label préféré « Smoking » et label alternatif « Fumeur ».

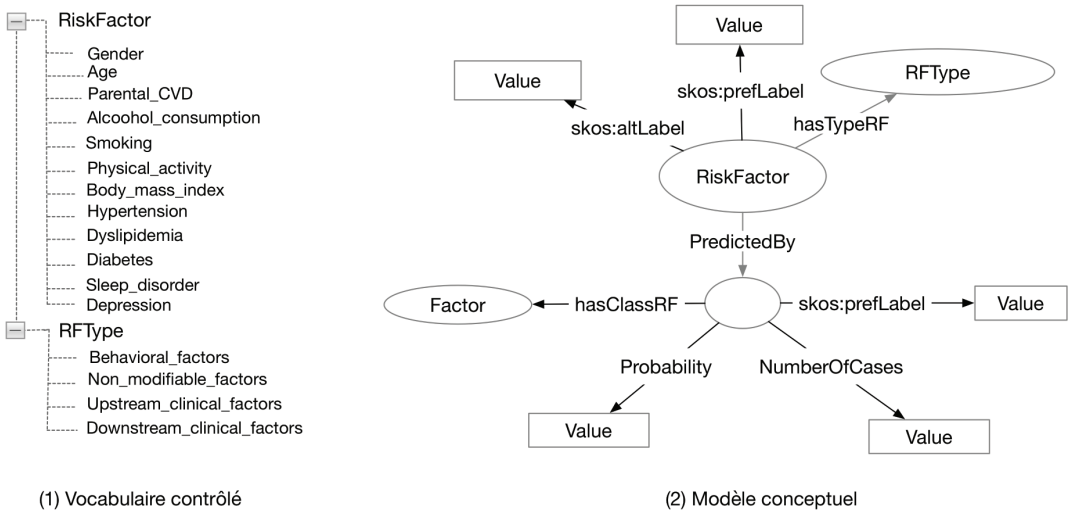

FIGURE *9 – Structure du vocabulaire contrôlé*

10. http ://bioportal.bioontology.org/ontologies/CARRE

— **Localisation de l'information** : le résultat obtenu à partir du tableau (A) est présenté Figure 10. Nous avons identifié à l'aide du vocabulaire contrôlé les termes «Smoking, predictive » dans le titre et « Obesity, Depression » dans le corps du tableau. Ainsi, le tableau (A) est annoté comme pertinent.

```
<P class="p72 ft1"><A href="#page_7">Table 4. </A>Risk of smoking according to predictive factors at baseline.</P>
<TABLE cellpadding=0 cellspacing=0 class="t10">
▪ ▪ ▪
<TR>
      <TD class="tr0 td61"><P class="p0 ft0"> </P></TD>
      <TD class="tr0 td54"><P class="p62 ft30">Obesity</P></TD>
      <TD class="tr0 td55"><P class="p62 ft31">67</P></TD>
      <TD class="tr0 td56"><P class="p81 ft31">1.25 <NOBR>(0.96–1.60)</NOBR></P></TD>
      <TD class="tr0 td57"><P class="p0 ft0"> </P></TD>
      <TD class="tr0 td58"><P class="p88 ft1">0.09</P></TD>
      <TD class="tr0 td56"><P class="p60 ft31">1.26 <NOBR>(0.97–1.62)</NOBR></P></TD>
      <TD class="tr0 td59"><P class="p62 ft31">0.08</P></TD>
      <TD class="tr0 td56"><P class="p81 ft31">1.27 <NOBR>(0.97–1.64)</NOBR></P></TD>
      <TD class="tr0 td25"><P class="p81 ft1">0.08</P></TD>
</TR>
▪ ▪ ▪
</TABLE>
```

FIGURE *10 – Exemple de tag réalisé sur un tableau à l'aide du vocabulaire contrôlé*

— **Extraction des tableaux pertinents** : une fois le tableau (A) déclaré pertinent, il est extrait sous forme d'un tableau associatif (voir Figure 11) « $numero\_de\_page \Rightarrow titre \Rightarrow contenu$ ».

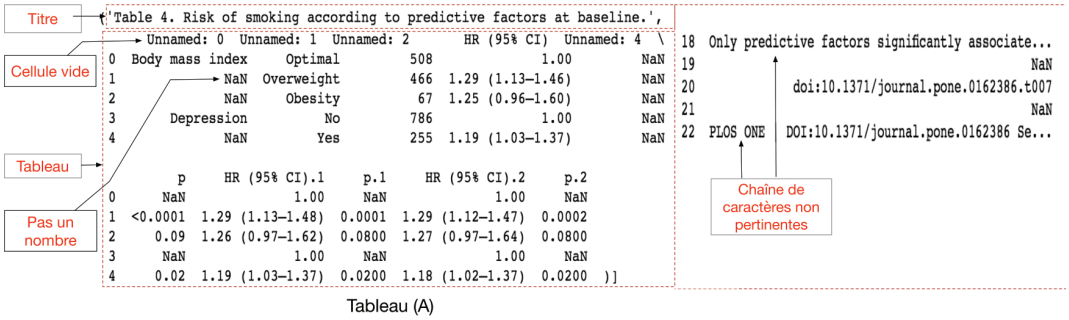

FIGURE *11 – Exemple de tableau extrait à partir du document HTML*

— **Extraction des colonnes pertinentes** : l'extraction des colonnes a été réalisée à l'aide du vocabulaire contrôlé et du modèle de triplet. Les colonnes extraites à partir du tableau (A) sont : (1) la probabilité décrite par l'en-tête portant l'étiquette « P » ; (2) le nombre de cas décrit par l'en-tête portant l'étiquette « No.of cases ».

('Table 4. Risk of smoking according to predictive factors at baseline.',

	Facteur	Label	NumberOfCases	Probability
1	Body_mass_index	Overweight	466	0.0002
2	Body_mass_index	Obesity	67	0.0800
3	Depression	Yes	255	0.0200

FIGURE *12 – Tableau (A) de la Figure 11 après nettoyage*

— **Nettoyage et mise en forme du tableau** : les résultats obtenus pour le tableau (A) sont présentés dans la Figure 12. Nous avons constaté que certains champs de la colonne «

probability » comporte des valeurs « NaN ». Cette valeur traduit la non-pertinence de la ligne. Nous avons appliqué un traitement de suppression de toutes les lignes contenant la valeur « NaN » dans la colonne « probability » pour l'ensemble des tableaux.

— **Construction de triplets** : le résultat obtenu pour le tableau (A) est présenté dans la Figure 13. Nous avons obtenu cinq nœuds blancs notés (_ :x_1, _ :x_2, _ :x_3). En prenant l'exemple du nœud blanc _ :x_1 associé à la première ligne du tableau, cinq triplets sont produits. Quatre triplets construits à partir de la ligne « 1 » et des colonnes « $Probability$ » et « $NumberOfCases$ » qui sont :
(_ :x_1,'$Class$','$Body\_mass\_index$'), (_ :x_1,'$Label$','$Overweight$'),
(_ :x_1,'$NumberOfCases$','466'), (_ :x_1,'$Probability$','0.0002'). Le cinquième triplet construit à partir du titre T et de la ligne « 1 » est ('$Smoking$','$PredictedBy$', _ :x_1).

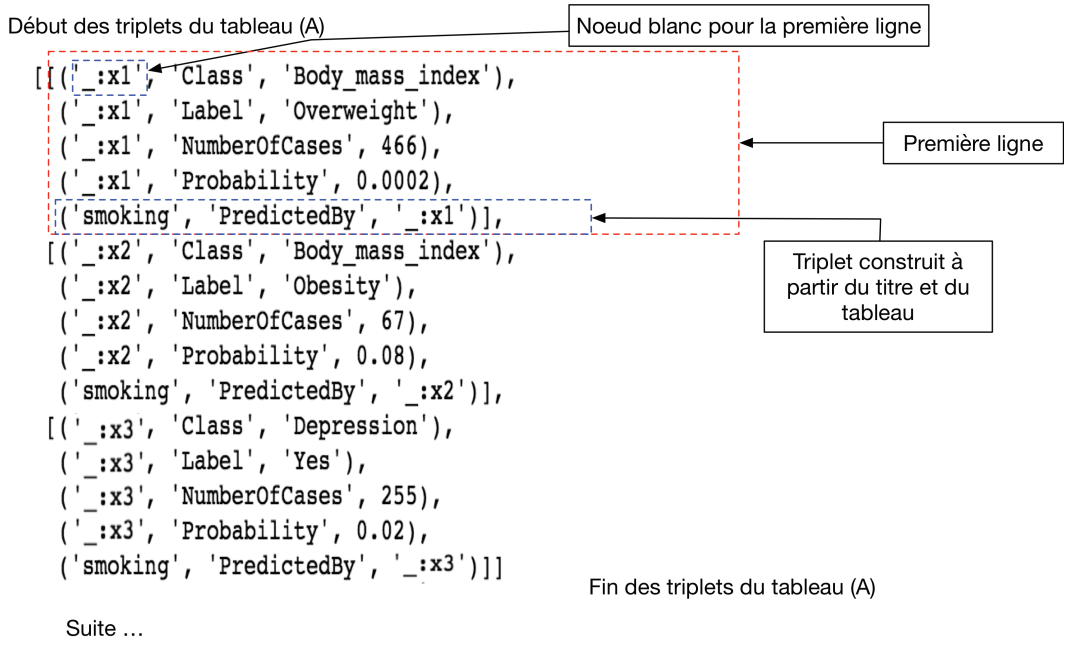

FIGURE *13 – Triplets obtenus à partir du tableau (A) de la Figure 11*

A l'issue de l'extraction des triplets, chaque facteur de risque est décrit selon le modèle conceptuel présenté dans la Figure 9-(2). Ce processus est réalisé en deux étapes : (1) analyser les triplets extraits ; (2) associer à chaque facteur de risque un ensemble de triplets.

Le résultat est stocké dans un Triple Store. Ainsi, le langage de requête SPARQL peut être utilisé directement pour interroger ce graphe. Outre son interrogation, le modèle obtenu est destiné à être utilisé dans une nouvelle approche d'évaluation du risque cardiovasculaire fondée sur la visualisation dynamique des interactions entre les facteurs de risque.

5 Evaluation

Nous avons montré dans la section précédente comment extraire et transformer des connaissances à partir des tableaux statistiques au format PDF. L'approche développée est adaptable à d'autres formats et à usage général. Elle est réalisée en deux étapes : (1) localisation et extraction de l'information dans les tables ; (2) élaboration du modèle conceptuel. Ces deux étapes sont fondées sur l'utilisation d'un vocabulaire contrôlé.

Afin de valider notre approche, nous avons conduit une expérimentation fondée sur l'interprétation par un expert de la sélection de tableaux statistiques correspondant à un sujet d'étude dans un document PDF. Pour évaluer l'approche d'extraction des tableaux, nous avons travaillé sur deux jeux de données disponibles au format PDF : le premier (D1) concerne le modèle statistique à l'origine de ce travail dans le domaine des maladies cardiovasculaires, la

démarche et les résultats obtenus sont présentés dans la section 4 ; le second (D2) concerne l'enquête internationale sur les transactions de change et de produits dérivés [11] dans le domaine financier. La Figure 14 présente l'exemple d'un tableau extrait du document et le résultat obtenu après extraction des triplets est présenté dans la Figure 15.

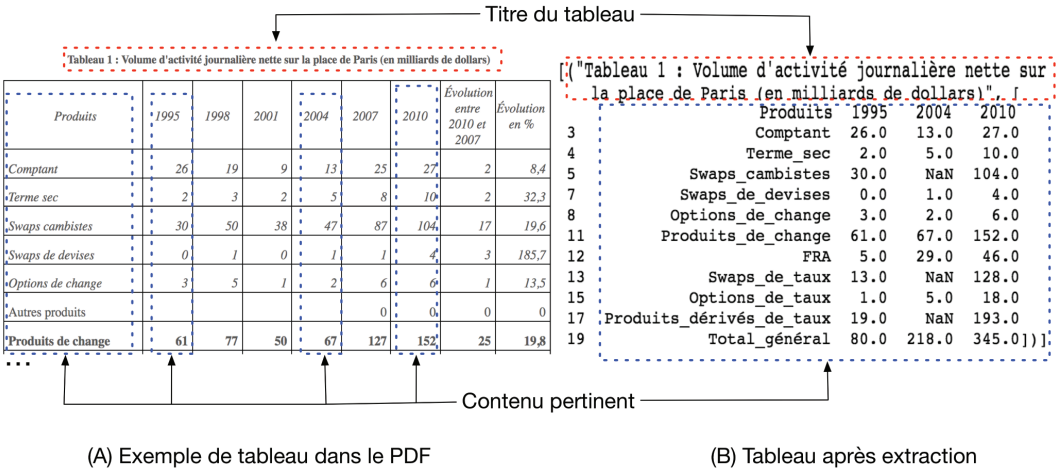

FIGURE *14 – Extraction d'un tableau figurant dans (D2)*

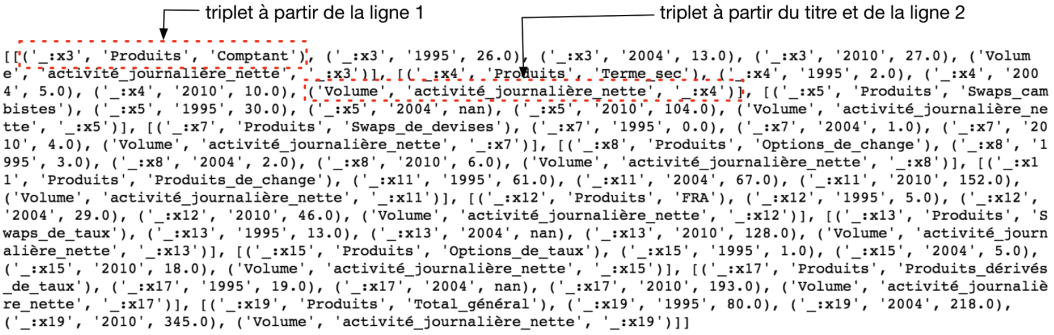

FIGURE *15 – Liste des triplets obtenus à partir du tableau de la Figure 14(B)*

Nous avons utilisé trois métriques pour évaluer les résultats de l'extraction sur (D1) et (D2) : la *précision*, le *rappel* et la *F-mesure*. Soit T_{fac} le nombre de tableaux identifié comme traitant des interactions entre les facteurs de risque cardiovasculaire (D1) et au volume d'activité de transactions (D2) ; la précision est le rapport entre T_{fac} et le nombre total de tableaux apparaissant dans chaque jeu de données ; le rappel est le rapport entre T_{fac} et les tableaux décrivant réellement les interactions entre les facteurs de risque cardiovasculaire (D1) et le volume d'activité de transactions (D2) ; la F-mesure est la moyenne quadratique combinant la précision et le rappel. Pour chaque jeu de données, nous avons exécuté une extraction avec et sans vocabulaire contrôlé. Une fois l'extraction réalisée, nous avons calculé la *précision*, le *rappel* et la *F-mesure*.

La Table 1 présente les résultats de l'extraction des tableaux avec les mesures de *précision*, *rappel* et *F-mesure* pour les jeux de données (D1) et (D2). Nous constatons que la précision

11. https ://www.banque-france.fr/sites/default/files/media/2016/11/24/enquete-triennale-principaux-resultats.pdf

Jeux de données	Vocabulaire	Précision	Rappel	F-mesure
D1	sans	0.40	0.90	0.55
	avec	**1.0**	**0.98**	**0.99**
D2	sans	0.50	0.45	0.47
	avec (1 terme)	**0.76**	**0.69**	**0.72**
	avec (2 termes)	**0.89**	**0.80**	**0.84**

TABLE *1 – Résultats expérimentaux*

de l'extraction donne de faibles résultats sur les deux jeux de données lorsqu'elle est réalisée sans le recours à un vocabulaire. Cette faible valeur de la précision est due au nombre élevé de tableaux extraits. La *précision* et la *F-mesure* sur (D1) augmentent lorsqu'un vocabulaire contrôlé est utilisé. Pour le jeu de données (D2), la *précision* et le *rappel* augmentent au fur et à mesure où le nombre de termes augmente dans le vocabulaire. Ces résultats indiquent que l'utilisation d'un vocabulaire adapté est déterminant pour optimiser l'extraction.

L'évaluation de cette approche sur le jeux de données (D1) fournit une performance presque parfaite pour le jeu de données (D1). Ce résultat s'explique principalement par : (1) l'approche développée sur le jeu de données (D1) ; (2) les tableaux décrivant les interactions entre les facteurs de risque cardiovasculaires dans le document HTML ont tous la même structure ; (3) le vocabulaire utilisé pour l'extraction est adapté au domaine. Sur le jeu de données (D2), la performance reste très satisfaisante. Nous prévoyons de tester l'approche d'extraction sur un volume plus important de jeux de données pour mieux évaluer l'approche et identifier des pistes d'amélioration.

Le cadre proposé dans cet article n'est pas limité à l'extraction de connaissances à partir d'études statistiques au format PDF, mais peut être appliqué à toutes ressources structurées sous forme de tableaux. L'originalité de cette approche est d'associer un modèle conceptuel aux tableaux figurant dans un document PDF.

Une autre expérimentation, en cours avec les chercheurs en statistiques, montre la difficulté de l'interprétation des connaissances représentées dans le modèle statistique. Les premiers résultats sont encourageants, ils démontrent outre un gain de temps, l'apport du langage SPARQL qui facilite l'accès aux connaissances (par exemple, filtrage sur la probabilité, le nom de facteur de risque, etc.).

6 Conclusion et perspectives

Dans cet article, nous avons décrit une méthode permettant la traduction d'un modèle statistique présenté sous forme de tableau et publié au format PDF, vers un modèle conceptuel représenté sous la forme d'un graphe. Nous avons apporté des solutions à deux problèmes dans le domaine de l'extraction de connaissances : (i) comment déterminer la pertinence des informations contenues dans des tableaux et sous quel format les extraire ; (ii) comment passer d'un format PDF non structuré à un format exploitable pour le traitement sémantique de l'information. Une réponse au second problème est constituée de la conversion d'un document PDF vers un format HTML respectant la structure des tableaux, puis l'extraction des informations pertinentes sous forme de triplets RDF. L'intérêt de cette approche est qu'elle permet d'extraire des connaissances implicites représentées dans des tableaux statistiques dans différents domaines.

Les résultats de nos premières expérimentations sur des ensembles de données de nature différentes sont encourageants, même s'ils doivent encore être améliorés. Plusieurs perspectives émergent comme l'ajout de l'exploitation du contenu complet du document (texte, figure, etc.). Le résultat obtenu est déjà intégré dans un système de visualisation [12] dynamique de connaissances appliqué aux interactions entre les facteurs de risque des maladies cardiovasculaires. En outre, l'approche est actuellement utilisée sur des études statistiques dans

12. http ://www-limics.smbh.univ-paris13.fr/MCVGraphViz/

le domaine des maladies cardiovasculaires et conduit à des modèles conceptuels différents. L'idée est de fusionner ces modèles en exploitant les connaissances expertes du domaine afin d'élaborer un modèle générique des interactions entre les facteurs de risque des maladies cardiovasculaires.

Références

CLARK C. A. & DIVVALA S. K. (2015). Looking beyond text : Extracting figures, tables and captions from computer science papers. In *Scholarly Big Data : AI Perspectives, Challenges, and Ideas, Papers from the 2015 AAAI Workshop, Austin, Texas, USA, January, 2015.*

CRESTAN E. & PANTEL P. (2010). Web-scale knowledge extraction from semi-structured tables. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, p. 1081–1082, New York, NY, USA : ACM.

ERMILOV I., AUER S. & STADLER C. (2013). User-driven semantic mapping of tabular data. In *Proceedings of the 9th International Conference on Semantic Systems*, I-SEMANTICS '13, p. 105–112, New York, NY, USA : ACM.

JUNYONG I. & SANGSEOK L. (2017). Statistical data presentation. *Korean Journal of Anesthesiology*, **70**, 267.

KLAMPFL S. & KERN R. (2015). Machine learning techniques for automatically extracting contextual information from scientific publications. In *Semantic Web Evaluation Challenges*, p. 105–116. Springer International Publishing.

LU W., ZHANG Z., LOU R., DAI H., YANG S. & WEI B. (2015). Mining rdf from tables in chinese encyclopedias. In *Proceedings of the 4th CCF Conference on Natural Language Processing and Chinese Computing - Volume 9362*, NLPCC 2015, p. 285–298, Berlin, Heidelberg : Springer-Verlag.

MENETON P., LEMOGNE C., HERQUELOT E., BONENFANT S., LARSON M.-G., VASAN R.-S., MÉNARD J., GOLDBERG M. & ZINS M. (2016). A global view of the relationships between the main behavioural and clinical cardiovascular risk factors in the gazel prospective cohort. *PLOS ONE*, **11**(9), 1–20.

MUÑOZ E., HOGAN A. & MILEO A. (2014). Using linked data to mine rdf from wikipedia's tables. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, p. 533–542, New York, NY, USA : ACM.

PIVK A., CIMIANO P., SURE Y., GAMS M., RAJKOVIČ V. & STUDER R. (2007). Transforming arbitrary tables into logical form with tartar. *Data Knowl. Eng.*, **60**(3), 567–595.

POWELL J. (2015). A librarian's guide to graphs, data and the semantic web. Chandos Information Professional Series, p. 268. Elsevier Science.

RIAZ A., TANVIRAND A. M. & MUHAMMAD A. Q. (2016). Information extraction from PDF sources based on rule-based system using integrated formats. In *Semantic Web Challenges*, p. 293–308. Springer International Publishing.

RONZANO F. & SAGGION H. (2016). Knowledge extraction and modeling from scientific publications. In A. GONZÁLEZ-BELTRÁN, F. OSBORNE & S. PERONI, Eds., *Semantics, Analytics, Visualization. Enhancing Scholarly Data*, p. 11–25, Cham : Springer International Publishing.

SAPORTA G. (2011). Probabilités, analyse des données et statistique. p. 622. 3ème édition révisée.

SHIGAROV A. O. (2015). Table understanding using a rule engine. *Expert Systems with Applications*, **42**(2), 929–937.

UNBEHAUEN J., HELLMANN S., AUER S. & STADLER C. (2012). *Knowledge Extraction from Structured Sources*, In S. CERI & M. BRAMBILLA, Eds., *Search Computing : Broadening Web Search*, p. 34–52. Springer Berlin Heidelberg : Berlin, Heidelberg.

WHO (2017). *World Health Statistics 2017 :Monitoring Health for the SDGs Sustainable Development Goals*. World Health Statistics Annual. World Health Organization.

WU J., KILLIAN J., YANG H., WILLIAMS K., CHOUDHURY S. R., TUAROB S., CARAGEA C. & GILES C. L. (2015). Pdfmef : A multi-entity knowledge extraction framework for scholarly documents and semantic search. In *Proceedings of the 8th International Conference on Knowledge Capture*, K-CAP 2015, p. 13 :1–13 :8, New York, NY, USA : ACM.

YEON-SEOK K. & KYONG-HO L. (2008). Extracting logical structures from HTML tables. *Computer Standards & Interfaces*, **30**(5), 296–308.