Apport du web sémantique pour améliorer la prévention des maladies cardiovasculaires

Rabia Azzi

▶ **To cite this version:**

Rabia Azzi. Apport du web sémantique pour améliorer la prévention des maladies cardiovasculaires. Web. Université Sorbonne Paris Cité, 2018. Français. NNT : 2018USPCD088 . tel-02872768

HAL Id: tel-02872768

https://theses.hal.science/tel-02872768

Submitted on 17 Jun 2020

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 13

École doctorale Galilée

N° attribué par la bibliothèque :
|__|__|__|__|__|__|__|__|__|__|

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PARIS 13

Discipline : Informatique

présentée et soutenue publiquement par

Rabia Azzi

le 28 novembre 2018

Apport du web sémantique pour améliorer la prévention des maladies cardiovasculaires

Directeur de thèse : Sylvie Despres, Professeur, Université de Paris 13
Co-directeur de thèse : Léopold Fezeu, MCF, Université de Paris 13
Co-encadrant de thèse : Jérôme Nobecourt, MCF, Université de Paris 13

Jury

Myriam Lamolle	Professeur, Université de Paris 8	Rapporteur
Gilles Venturini	Professeur, Université de Tours	Rapporteur
Sylvie Despres	Professeur, Université de Paris 13	Directeur de thèse
Léopold Fezeu	MCF, Université de Paris 13	Co-directeur de thèse
Jérôme Nobécourt	MCF, Université de Paris 13	Co-encadrant
Juliette Dibie-Barthélemy	Professeur, AgroParisTech	Présidente du jury
Pierre MENETON	Chercheur, INSERM	Examinateur
Gayo Diallo	MCF, Université de Bordeaux	Examinateur
Rosy Tsopra	Médecin, Dr Informatique médicale	Invitée

Thèse préparée au Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé (LIMICS)
UMR_S 1142, 74 Rue de Marcel Cachin, 93017 Bobigny, France

THESE

À mes parents, Juliette et Aylan

Remerciements

Faire un doctorat c'est apprendre à devenir chercheur. Mais avant d'être une recherche scientifique, c'est avant tout une recherche de soi. On se découvre des forces et des faiblesses. C'est un chemin long et semé d'embûches qui ne peut se réaliser sans un soutien important.

J'exprime tout d'abord mes remerciements et ma gratitude à Sylvie Despres, mon directeur de thèse, pour sa compétence, sa pédagogie et surtout sa patience. Nous nous sommes rencontrés lors de mon master 1 et tu as su me transmettre ton expérience et ta passion pour l'informatique. Ces trois années ont été riches en enseignements grâce à toi, tant au niveau personnel que professionnel.

Mes remerciements vont aussi à Léopold Fezeu, mon co-directeur de thèse, pour son expertise sur les maladies cardiovasculaires qui m'a permis d'avoir une vision plus claire des problématiques du sujet. Merci également à Jérôme Nobécourt de m'avoir encadré et notamment de m'avoir apporté son expertise dans le cadre de la thèse et de l'enseignement.

Je tiens également à exprimer mes remerciements à Myriam lamolle et Gilles Venturini qui ont accepté d'être les rapporteurs de cette thèse et pour le temps qu'ils auront bien voulu y consacrer. Meri également à Juliette Dibie-Barthélemy, Pierre Meneton ainsi qu'à Gayo Diallo qui ont accepté d'être examinateur.

Merci à tous les autres membres du LIMICS que j'ai eu la chance de côtoyer. Merci à tous les amis pour les discussions et les bons moments passés ensemble : Rosy, Rahma, Bernard, Gilles, Karima, Mélanie, etc. Comme Rosy l'a dit « vive le bureau des confidences ... ! ». Je remercie Rosy pour son soutien, son encouragement et les échanges que nous avons eus autour de mon sujet de thèse. Merci également à Rahma pour ses relectures rigoureuses.

Merci à Brigitte ma belle-mère, qui m'a accueillie comme son propre fils. Merci de m'avoir toujours encouragé.

Je souhaite remercier spécialement Juliette pour son soutien et sa patience tout au long de cette thèse.

Enfin, je remercie ma famille, malgré la distance, vous êtes pour moi une source d'énergie et de courage.

Sommaire

Sommaire ..... **vii**

Préambule **1**

Introduction **3**

Chapitre 1 Contexte applicatif et scientifique **9**

Chapitre 2 Exploitation des présentations de résultats statistiques **39**

Chapitre 3 Modélisation des connaissances **65**

Chapitre 4 MCVGraphViz, un outil de visualisation **77**

Chapitre 5 Systèmes de recommandation **103**

Chapitre 6 Qualification nutritionnelle des recettes de cuisine **127**

Conclusion et perspectives **155**

Bibliographie **186**

Chapitre A Evolution de l'interface de l'outil MCVGraphViz **187**

Chapitre B Qualification nutritionnelle des recettes **191**

Table des matières **195**

Index des tableaux **197**

Index des figures **201**

Préambule

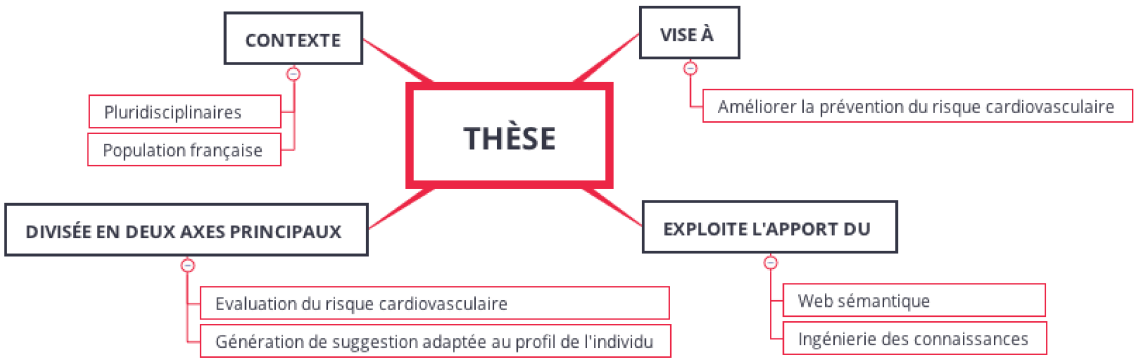


FIGURE 1 – Caractéristiques générales des travaux de la thèse

Chapitre	Problématique	Objectif	Résultat
1	Prévenir le risque cardiovasculaire	Exploiter les apports du web sémantique et de l'ingénierie des connaissances	Un écosystème de prévention du risque cardiovasculaire
2	Exploiter les connaissances issues des résultats des études statistiques publiées sous divers formats	Extraire les connaissances implicites et réduire les erreurs dues à l'interprétation des relations dans les tableaux statistiques	Un modèle conceptuel du domaine sous forme d'un graphe
3	Comparer différents modèles conceptuels obtenus à partir de différentes études statistiques	Élaborer un modèle générique	Un modèle générique du domaine
4	Exploiter les techniques de visualisation pour améliorer la compréhension des liens entre les différents facteurs de risque cardiovasculaire	Mettre en place une approche pour prendre en compte l'interaction entre les facteurs de risque	Un outil de visualisation dynamique de connaissances
5	Construire un système de recommandation à partir de sources de connaissances hétérogènes	Proposer un système de recommandation destiné à la prévention du risque cardiovasculaire	Un système de recommandation modulaire pour la prévention du risque cardiovasculaire
6	Apparier différentes ressources textuelles et des données structurées pour améliorer le calcul des valeurs nutritionnelles des recettes de cuisine	Construire un module de qualification nutritionnelle des recettes de cuisine qui va s'intégrer dans l'écosystème de prévention du risque cardiovasculaire et mieux renseigner et guider les utilisateurs dans l'adoption d'un mode de vie sain	Une approche de qualification nutritionnelle des recettes de cuisine et réalisation d'un prototype

TABLE 1 – Tableau synoptique du document

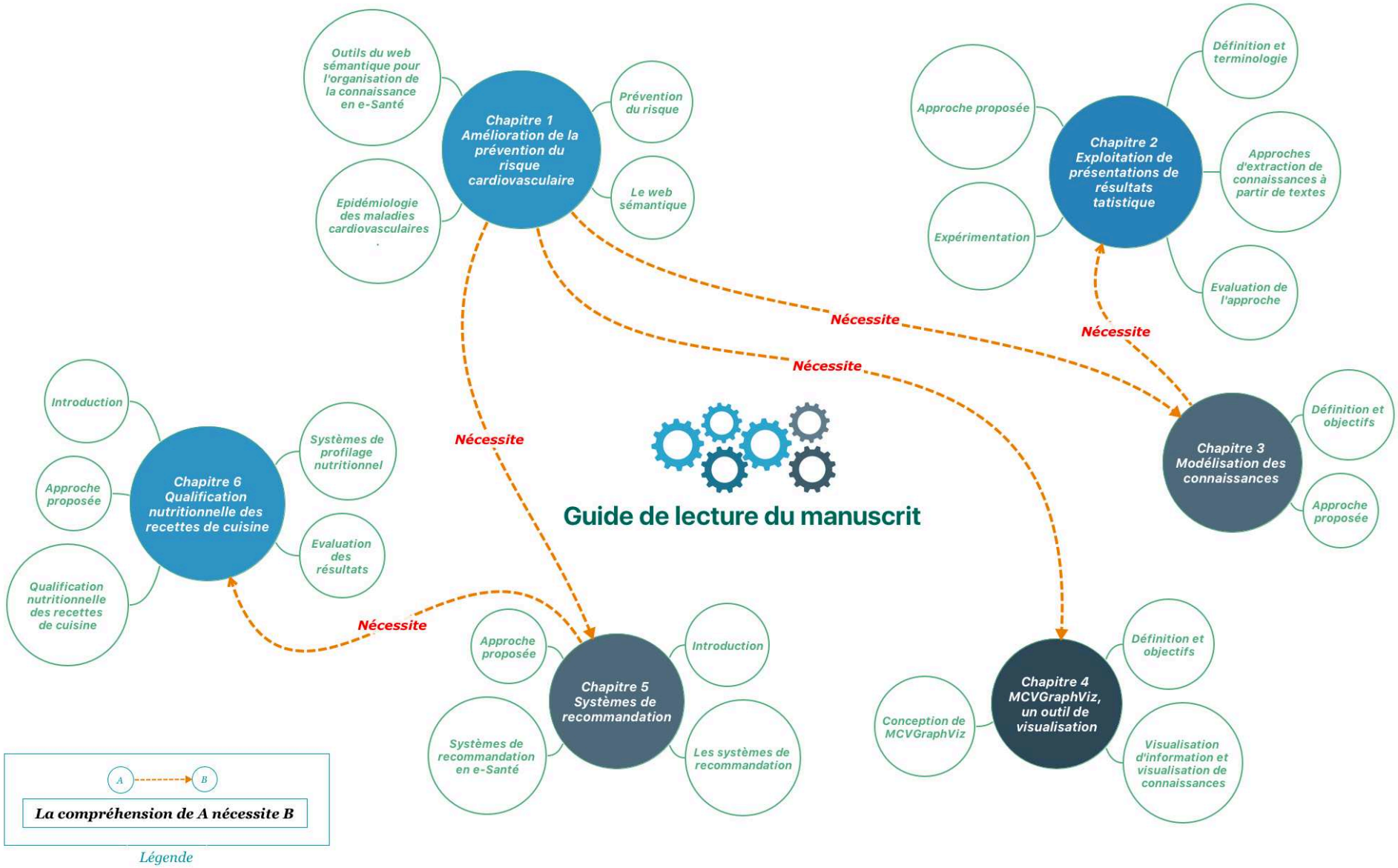


FIGURE 2 – Guide de lecture du manuscrit

Introduction

Les technologies de e-santé ont ouvert de nouveaux horizons pour promouvoir une autonomie accrue des patients. Bien que les technologies de l'information soient maintenant largement utilisées, il existe souvent une disparité entre les connaissances scientifiques et technologiques sous-jacentes aux pratiques de soins de santé et les représentations cognitives de la maladie. Cela s'explique par l'avènement de nouvelles technologies embarquées, l'utilisation généralisée de dispositifs à déploiement universel tels que les smartphones et les tablettes, ainsi que par les progrès des communications sans fil. Ces dernières années, l'explosion du nombre de sources documentaires sur le web rendant parfois la recherche d'information laborieuse, d'autant plus lorsqu'il s'agit de retrouver des informations fiables et pertinentes. Ainsi, les informations sont souvent mal exploitées à cause de leur hétérogénéité et de la difficulté pour les localiser. En effet, la nature des documents publiés présente une grande variabilité. Pour optimiser l'exploitation de ces informations, de nouveaux outils sont nécessaires pour structurer et rendre interopérable les différentes ressources.

L'ambition de cette thèse est de construire un écosystème apprenant pour la prévention du risque cardiovasculaire. Il s'agit d'un ensemble composé d'entités (modules, outils, etc.) et de leurs utilisateurs (patient, médecin, aidant, etc.) interagissant dans un environnement technologique. Cet environnement est constitué d'un ensemble de services informatiques utilisant les technologies du web. Les différents acteurs s'engagent à contribuer à la réussite des objectifs de cet écosystème qui est conçu pour faciliter les échanges et le partage de ressources entre les différents acteurs impliqués. Il est organisé autour d'un ensemble d'espaces de collaboration dédiés aux utilisateurs travaillant collectivement sur un même problème. Il permet ainsi à l'utilisateur de construire et de partager de nouvelles connaissances à partir des interactions avec les différents utilisateurs et les ressources mises à disposition. Les informations contenues dans ces ressources sont relatives à : (i) l'évaluation du risque cardiovasculaire ; (ii) la prise en charge selon la nature du risque ; (ii) la description des processus de prise de décision qui associe à une situation particulière une recommandation dépendante du contexte.

La prévention de risque cardiovasculaire est un processus complexe comportant des facteurs de risque pouvant interagir du niveau moléculaire au niveau sociétal et environnemental. Des méthodes d'analyse des données adéquates sont nécessaires pour améliorer la compréhension de cette complexité. Leur finalité est l'obtention d'évaluations probantes de haute qualité à des fins d'information des différents utilisateurs (patients, professionnels de santé) et l'aide à la mise en place d'interventions adaptées. Toutefois, la plupart du temps, ces connaissances sont issues d'études épidémiologiques réalisées par des experts et publiées sur des supports statiques destinés à des êtres

humains et incompréhensibles par des machines. De plus, la clé de la réussite d'une intervention est le changement du comportement des utilisateurs ce qui a conduit au développement de systèmes éducatifs personnalisés. L'opérationnalisation de ces systèmes personnalisés nécessite la formalisation des connaissances du domaine afin de construire des suggestions adaptées aux différents types d'utilisateurs.

Contexte applicatif. Les travaux présentés dans cette thèse s'inscrivent dans une démarche interdisciplinaire. Ils ont été menés dans le contexte d'une collaboration entre le Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS [1]), Université Paris 13 et l'Equipe de Recherche en Epidémiologie Nutritionnelle (EREN [2]), Université Paris 13.

Les maladies cardiovasculaires (MCVs) constituent la première cause de mortalité et de morbidité en France chez les hommes et les femmes engendrant des coûts annuels importants. La prévention de ces maladies constitue avec leur dépistage précoce et leur prise en charge rapide et efficace un moyen possible de réduire ce coût. Cette prévention passe par l'identification des facteurs de risque modifiables associés à ces maladies chroniques. Ainsi, plusieurs plans d'action (prévention et lutte) sont mis en œuvre contre ces maladies non transmissibles (Assurance-Maladie, 2010).

Ces dernières années, il a été observé à l'échelle mondiale, une reconnaissance du potentiel des technologies informatiques pour leur contribution à l'amélioration de la santé publique. En 2018, l'assemblée mondiale de la santé a approuvé le Plan d'action pour la mise en œuvre de la stratégie mondiale de lutte contre les maladies non transmissibles (WHO, 2018). A l'échelle internationale, on assiste à une prise de conscience accrue (i) des défis posés par les maladies non transmissibles et (ii) des éléments de preuve à l'appui des interventions efficaces.

Parallèlement à la création d'alliances (notamment la participation des agences internationales, de la communauté scientifique et de la santé publique et des Organisations Non Gouvernementales (ONG) à la sensibilisation, à la recherche et à la collaboration), plusieurs rapports importants ont été rédigés. En Europe, une attention nouvelle a été accordée aux principaux facteurs de risque des maladies non transmissibles. L'élaboration de la nouvelle politique européenne de la santé, Santé 2020, permet à la région européenne de l'OMS d'ouvrir un nouveau chapitre dans ce domaine et un engagement renouvelé pour la santé publique. En France, des organismes et des programmes de recommandations sont mis en place tels que : (i) la Haute Autorité de santé (HAS [3]) qui propose des recommandations conçues pour aider le praticien et le patient à déterminer les soins les plus appropriés à des situations cliniques ; (ii) le Programme National Nutrition Santé (PNNS [4]) qui a pour objectif de contribuer à réduire l'incidence de ces maladies en contribuant à l'éducation des Français sous forme d'objectifs relatifs à l'alimentation et à l'activité physique.

Dans de nombreux domaines, et en particulier en santé, l'accès aux informations et leur utilisation sont des enjeux majeurs pour le grand public et les professionnels. Depuis son apparition au début des années 1990, le web a profondément transformé la société. Il est désormais omniprésent dans nos vies, que ce soit dans notre façon de communiquer, de travailler, etc. Ainsi, l'information médicale devient de plus en plus disponible et accessible. De nombreux sites existent pour informer les praticiens, optimiser leurs pratiques et leurs organisations, d'autres permettent aux usagers d'accéder

1. http://www.limics.fr/
2. https://eren.univ-paris13.fr/index.php/fr/
3. https://www.has-sante.fr
4. http://www.mangerbouger.fr/PNNS

à des informations pas toujours certifiées. Les sites de référencement des publications médicales telles que PubMed (NCBI, 2018) et CISMeF (CISMeF, 2018) vont dans ce sens. A titre d'exemple, PubMed indexe plus de 2275544 références bibliographiques sur les maladies cardiovasculaires et une recherche à partir de CISMeF, donne 6539 résultats sur les maladies cardiovasculaires.

Le contexte de la thèse est celui de l'ingénierie des ontologies et du web sémantique appliqués à la prévention et au suivi nutritionnel de patients présentant des risques ou souffrant de maladies cardiovasculaires. L'objectif du travail est la construction d'un service capable d'évaluer le risque cardiovasculaire d'un individu et de fournir des suggestions personnalisées fondées sur le profil de l'internaute ou du patient. Cette approche nécessite la mise en place d'une démarche interdisciplinaire faisant collaborer des chercheurs dans les domaines de l'informatique, de l'épidémiologie et de la nutrition. L'importance de cette collaboration se justifie par le besoin de produire des recommandations étayées par des recherches attestées dans le domaine de la nutrition et de la santé.

La modélisation des connaissances sur les facteurs de risque cardiovasculaire fait appel à l'ingénierie des connaissances, qui vise à proposer des concepts, des méthodes et des techniques permettant de modéliser, de formaliser, d'acquérir des connaissances dans les organisations dans un but d'opérationnalisation, de structuration ou de gestion au sens large (par exemple, valider les modèles qualitatifs conçus lors de l'étape précédente en conservant une traçabilité entre les modèles au niveau des connaissances et les modèles pour l'implémentation.) (Teulier *et al.*, 2005). Lors du passage de l'opérationnalisation à la validation, le processus de validation doit être conçu pour prendre en compte les retours des experts et des utilisateurs afin d'améliorer les modèles qualitatifs élaborés et de progresser dans l'analyse du problème et la définition des méthodes élaborées pour le résoudre (Charlet *et al.*, 2000).

Actuellement, le cadre technique de la formalisation des connaissances est fourni par le web sémantique. Le web sémantique est un projet dont le but est d'améliorer la présentation des documents en ligne afin de les rendre compréhensibles par des processus automatiques, permettant ainsi à des machines de chercher, de combiner, de gérer et de produire des connaissanes. Sous l'égide du W3C, des standards sont apparus pour exprimer de manière formelle les connaissances, tels que RDF (Klyne et Carrol, 2004) et OWL (Motik *et al.*, 2012), et pour les interroger, tel que le langage de requêtes SPARQL (Prud'hommeaux et Seaborne, 2008), GraphQL (GraphQL, 2018), HyperGraphQL (HyperGraphQL, 2018). Le web sémantique a déjà gagné en popularité en tant que plateforme d'intégration et d'analyse de données dans le domaine de la santé. De nombreux référentiels et des ensembles de données disponibles publiés selon les meilleurs principes et pratiques du web sémantique sont disponibles. Les interfaces homme-machine pour interroger, afficher et visualiser les données sémantiques sur la santé permettant l'exploration de données pour la découverte de connaissances cachées dans les données de santé.

Problématique. Actuellement, on assiste à une explosion des données à la fois en volume et en diversité. Mais la plupart des ensembles de données disponibles dans le domaine de la santé continuent à être produits dans des formats propriétaires allant de ceux qui sont pratiques, précis, indexés et bien gérés à ceux qui sont incomplets, incohérents, potentiellement inexacts et extrêmement volumineux. Par exemple, (WHO, 2017) continue à publier des rapports sur les problèmes de santé mondiaux et à fournir un accès à d'énormes données, des analyses statistiques et des recommandations sur les maladies non transmissibles. De nombreux autres établissements de santé tels que Pub-

Med, CISMeF, HAS contiennent des liens sélectifs vers différents référentiels de bases de données de santé. Cependant, la plupart des données continuent à être publiées au format PDF ou sous forme de feuilles de calcul, rendant difficile leur traitement ultérieur y compris pour le processus d'agrégation et d'interconnexion de données. L'informatisation de la santé, telle que l'utilisation quotidienne des systèmes d'information et des dispositifs médicaux technologiques dans les hôpitaux et autres institutions médicales, a déjà produit et continuera de produire une quantité considérable de données issues des dossiers médicaux, de la surveillance des patients, etc. Ces dernières années, de nombreuses tentatives d'utilisation du web sémantique pour réduire la complexité du traitement des données et résoudre certains problèmes d'intégration classiques sont mises en œuvre.

Afin de répondre aux attentes afférentes à la productivité accrue du volume des données qui découle de l'utilisation des technologies du web sémantique dans le domaine des soins de santé, un certain nombre de problèmes doivent être résolus :

— Outre le problème d'intégration des données, l'interaction de l'utilisateur avec les données constitue un autre défi. La difficulté réside dans la gestion des recherches, la navigation dans les données et la présentation des données. La difficulté est de mettre au point des outils permettant de trouver efficacement des ressources ou des données, en particulier s'il peut y avoir un ensemble de liens et d'interconnexions possibles entre les données publiées provenant de nombreuses sources différentes. De telles associations pourraient être utilisées en amont pour fournir de nouveaux contextes et de nouvelles possibilités de résultats de recherche plus efficaces (modélisation des connaissances d'un domaine). En outre, une autre exigence importante consiste à fournir une présentation cohérente de ces données et à faciliter la navigation et l'exploration efficace de ces données et cela à partir de différents périphériques.

— Un autre défi consiste à savoir comment utiliser ces énormes quantités de données pour construire de nouveaux modèles et transformer ces données en un ensemble de connaissances susceptible d'améliorer les pratiques en e-santé. Par exemple, une application d'exploration de données pourrait explorer les données des patients (tels que les habitudes alimentaires, l'activité physique, ou les antécédents familiaux, etc.) pour détecter les causes des maladies et proposer une recommandation adaptée à moindre coût. Ces dernières années, le web sémantique gagne du terrain pour relever ces défis.

— Plusieurs ressources peuvent décrire des notions utilisées dans un même domaine selon des points de vue différents. Il est difficile de parvenir à un accord commun et large pour décrire un domaine spécifique, aussi des efforts sont nécessaires pour standardiser les vocabulaires dans les domaines de la santé. Les données continueront à être décrites en utilisant différents vocabulaires, mais pour intégrer ces informations dans une vue cohérente, diverses stratégies d'appariement et d'alignement sont nécessaires. Il existe cependant peu d'interfaces conviviales pour interroger, parcourir, afficher et visualiser les données sémantiques de manière significative. Le développement d'interfaces efficaces améliore considérablement l'adoption du web sémantique dans le domaine de la santé.

Le but de cette thèse est par conséquent de définir des méthodes pour faciliter l'accès à des ressources du domaine de la santé (HAS, PNNS, LANGUAL, etc.), de construire des outils et des stratégies pour automatiser les processus de qualification des ressources, d'évaluer le risque par une approche de visualisation dynamique des

interactions entre les facteurs de risques. Les approches et les prototypes développés au cours de cette thèse se veulent les plus génériques possible. Par exemple, dans le cas de l'évaluation du risque, l'approche est applicable à n'importe quel domaine de la gestion du risque (nucléaire, ferroviaire, santé, etc.).

Les contributions de cette thèse sont : (i) un système d'évaluation du risque cardiovasculaire qui prend en compte les interactions entre les facteurs de risque ; (ii) un outil de visualisation de connaissances représenté sous forme d'un graphe ; (iii) une approche de qualification nutritionnelle des recettes de cuisine, implémentée sous forme d'un service web ; (iv) un écosystème apprenant avec une architecture modulaire composée de services web. Cette approche facilite son évolution dans le domaine des maladies cardiovasculaires et sa réutilisation dans d'autres domaines.

Organisation du mémoire. La premier chapitre présente le contexte de ce travail qui s'inscrit dans le cadre de proposition d'un écosystème apprenant pour améliorer la prévention du risque cardiovasculaire. La section 1.1 introduit les objectifs de cet écosystème qui s'appuient notamment sur les connaissances issues de l'épidémiologie des maladies cardiovasculaires (section 1.2) et de la prévention du risque dans différents domaines (section 1.3). Le contexte technique de cet écosystème, présenté dans la section 1.4, est fourni par le web sémantique. Plusieurs de ces applications dans le domaine de la e-santé sont présentées dans la section 1.5.

La deuxième chapitre vise à fournir une approche d'extraction automatiquement des connaissances à partir de textes. Dans notre étude, nous nous intéressons plus précisément à l'extraction et l'interprétation de connaissances à partir de tableaux statistiques publiés au format PDF. Nous commençons par décrire la démarche statistique d'interprétation des résultats et les méthodes visuelles les plus utilisées pour les présenter. Ensuite, nous présentons et comparons les approches d'extraction des connaissances à partir de ces présentations avant d'exposer notre approche. Nous expérimentons ensuite l'approche d'extraction proposée sur deux jeux de données relatives aux domaines des maladies cardiovasculaires et de la finance. Enfin, l'approche est évaluée.

Dans le troisième chapitre, nous présentons un modèle conceptuel générique décrivant les facteurs de risque et leurs interactions. Nous abordons la problématique de modélisation conceptuelle des connaissances en nous fondant sur les concepts de l'ingénierie des connaissances afin d'offrir une modélisation consensuelle et partagée sur les maladies cardiovasculaires. Nous détaillons et discutons la démarche adoptée afin de démontrer l'utilité d'un modèle conceptuel générique pour un domaine.

L'objet du quatrième chapitre est de présenter l'outil MCVGraphViz proposé pour la visualiser le graphe des interactions entre facteurs de risque. Il s'appuie sur le modèle conceptuel construit pour offrir à l'utilisateur une interface de visualisation permettant d'accéder de façon progressive et dynamique au contenu du graphe. De plus, le système collecte les actions effectuées au cours de la navigation permettant de constituer une trace et de la stocker sous la forme d'un chemin (nœud de départ, nœud de destination).

Le système de recommandation est le sujet du cinquième chapitre. Nous commençons par nous intéresser aux systèmes de recommandation et leur utilisation dans le domaine de la santé. Nous nous intéressons à la fois aux aspects techniques et à la manière dont les utilisateurs perçoivent un tel système. Puis nous proposons une approche de recommandation modulaire adaptée à l'utilisateur en nous basant sur leurs besoins et un ensemble de critères identifiés dans la littérature. Après avoir abordé le choix de la modularité, nous présenterons les interfaces implémentant chaque module alliant ainsi l'expertise humaine à l'efficacité informatique pour améliorer la prévention du risque cardiovasculaire.

L'objet du sixième chapitre est la présentation du problème de la qualification nutritionnelle des recettes de cuisine qui consiste à automatiser l'attribution d'un score nutritionnel à une recette et son implémentation. L'objectif du travail présenté dans ce chapitre est de construire un module de qualification nutritionnelle des recettes qui s'intègre dans l'écosystème de prévention du risque cardiovasculaire, pour fournir aux utilisateurs des contenus et des outils pour mieux les renseigner et les guider dans l'adoption d'un mode de vie sain. L'approche adoptée se décompose en quatre étapes : enrichissement lexical en annotant les termes désignant les produits composant les ingrédients ; génération du fichier de calcul à partir d'un patron lexical et interrogation de la table de composition ; calcul et attribution du score ; traduction du score sur une échelle graphique. Cette méthodologie est fondée sur l'appariement entre des ressources textuelles (corpus de recettes) et des données structurées (table de composition Nutrinet). La ressource termino-ontologique MIAM est utilisée pour améliorer la qualité de l'appariement et permettre la caractérisation nutritionnelle des recettes.

Chapitre 1

Contexte applicatif et scientifique

Sommaire

1.1 Projet d'amélioration de la prévention du risque cardiovasculaire **10**
1.2 Epidémiologie des maladies cardiovasculaires **10**
1.2.1 Généralités sur les maladies cardiovasculaires 10
1.2.2 Statistiques des maladies cardiovasculaires 10
1.2.3 Les facteurs associés aux maladies cardiovasculaires 11
1.3 Prévention du risque **14**
1.3.1 Définition du risque 14
1.3.2 Définition de la prévention 15
1.3.3 Systèmes d'évaluation du risque 16
1.3.4 Approches de prévention sensibles au domaine 20
1.3.5 Vers un système de prévention du risque cardiovasculaire informatisé 21
1.4 Le web sémantique **21**
1.4.1 Présentation générale du web sémantique 21
1.4.2 Les standards du W3C 24
1.5 Outils du web sémantique pour l'organisation de la connaissance en e-santé **30**
1.5.1 Thésaurus et ontologies 31
1.5.2 Référentiels de données sémantiques 36
1.6 Conclusion partielle **38**

Ce chapitre présente le contexte de ce travail qui s'inscrit dans le cadre de proposition d'un écosystème apprenant pour améliorer la prévention du risque cardiovasculaire. La section 1.1 introduit les objectifs de cet écosystème qui s'appuient notamment sur les connaissances issues de l'épidémiologie des maladies cardiovasculaires (section 1.2) et de la prévention du risque dans différents domaines (section 1.3). Le contexte technique de cet écosystème, présenté dans la section 1.4, est fourni par le web sémantique. Plusieurs de ces applications dans le domaine de la e-santé sont présentées dans la section 1.5.

Texte de référence :

S. Adriouc, R. Azzi et L. Fezeu. (2017). Des preuves convaincantes de l'association entre nutrition et maladies cardiométaboliques, Dossier : Rôle de la nutrition dans les maladies cardiométaboliques. Serge Hercberg, La revue du praticien, Mars 2017, pp 13-18.

1.1 Projet d'amélioration de la prévention du risque cardiovasculaire

L'objectif est de proposer des méthodes et de créer des outils logiciels pour l'aide à la prévention du risque cardiovasculaire et, plus généralement, pour la gestion des connaissances de ce domaine. Les connaissances étudiées sont issues des institutions nationales et internationales sous forme de documents. Elles sont composés de l'ensemble des recommandations nécessaires à l'évaluation et à la prise en charge des individus à risque ou souffrants de maladie cardiovasculaire. L'objectif est de formaliser ces connaissances de manière à les rendre utilisables par des systèmes intelligents en fournissant les outils pour leur maintenance et leur mise à jour.

1.2 Epidémiologie des maladies cardiovasculaires

1.2.1 Généralités sur les maladies cardiovasculaires

Les travaux concernant la compréhension des causes des maladies cardiovasculaires constituent un domaine relativement récent en médecine. En fait, les maladies cardiovasculaires étaient relativement rares avant le début des années 1900 (Teufel, 2018), mais avec l'arrivée des antibiotiques et les progrès de la science, les décès prématurés dus à des maladies infectieuses ont chuté. En outre, le développement des transports et des industries alimentaires a réduit le besoin de se procurer nos propres aliments grâce à l'agriculture et donc à des modes de vie plus sédentaires. Cette combinaison de facteurs entraîne une augmentation rapide des maladies cardiaques (voir figure 1.1).

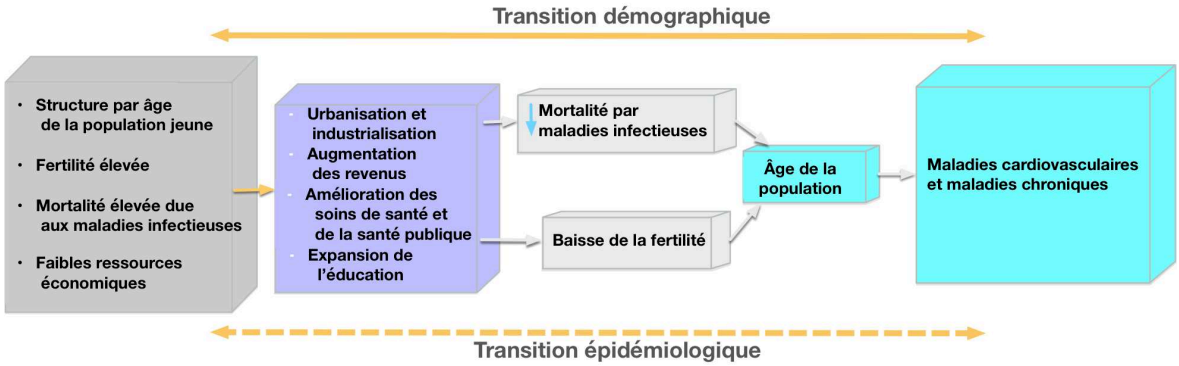


Figure 1.1 – Transition démographique des maladies chroniques

Les maladies cardiovasculaires constituent un ensemble de troubles affectant le cœur et les vaisseaux sanguins. Les infarctus du myocarde et les accidents vasculaires cérébraux (AVC) sont généralement des événements aigus et principalement dus à l'obstruction d'une artère empêchant le sang de parvenir au cœur ou au cerveau. Leur cause la plus courante est la constitution d'une plaque d'athérome sur les parois internes des vaisseaux sanguins alimentant ces organes. Les AVC peuvent aussi résulter du saignement d'un vaisseau sanguin cérébral (WHO, 2007).

1.2.2 Statistiques des maladies cardiovasculaires

Selon l'OMS (WHO, 2017) les maladies cardiovasculaires sont la première cause de mortalité dans le monde. L'OMS estime en 2015 à 17,7 millions le nombre de décès

imputables aux maladies cardiovasculaires, soit 31% de la mortalité mondiale totale. Parmi ces décès, 7,4 millions sont dus à une cardiopathie ischémique et 6,7 millions à un AVC.

Les maladies cardiovasculaires sont également la première cause de décès prématuré à travers le monde : 42% de l'ensemble des décès survenant avant l'âge de 75 ans leur sont imputables. Ceci souligne l'urgence de promouvoir la prévention de ces pathologies en agissant sur leurs facteurs de risque modifiables.

En France, les maladies cardiovasculaires sont responsables de 29% des décès (150 000 par an), juste derrière les cancers qui représentent 30% de la mortalité. Leur coût direct est estimé à 13 millions d'euros et leur coût total à 22 millions d'euros (Assurance-Maladie, 2010). Elles occupent le premier rang des Affections de Longue Durée (ALD) avec 2,8 millions de personnes concernées et concentrent 27% des dépenses consacrées aux ALD, soit 18,7 milliards d'euros de remboursements.

1.2.3 Les facteurs associés aux maladies cardiovasculaires

Un facteur est associé à une maladie cardiovasculaire si sa présence ou son absence modifie le risque de développer cette maladie. On distingue les facteurs de risque et les facteurs protecteurs. Un facteur de risque est une condition ou une habitude qui rend une personne plus susceptible de développer une maladie. Un facteur protecteur est une condition ou habitude dont la présence diminue le risque de développer une maladie.

La recherche de facteurs associés à une maladie cardiovasculaire est fondamentale. Elle permet : (i) d'identifier les sujets à risque d'événement cardiovasculaire ; (ii) de définir les stratégies d'intervention en fonction des facteurs de risque présents. Plusieurs facteurs de risque (voir tableau 1.1), responsables de l'augmentation d'incidence des maladies cardiovasculaires, ont été identifiés grâce aux études de cohorte et aux essais thérapeutiques randomisés. Le tableau 1.1 présente un résumé des principaux facteurs de risque cardiovasculaire (Gaye *et al.*, 2016 ; Viot-Blanc, 2010 ; Piepoli *et al.*, 2016). Nous pouvons les classer en deux types de facteurs de risque : (i) les facteurs de risque non modifiables : l'âge, le sexe, ou les facteurs génétiques ; (ii) les facteurs de risque dits modifiables, car accessibles à une intervention : le tabagisme, la sédentarité, la mauvaise alimentation et l'usage nocifs de l'alcool, etc.

Face à la morbidité importante liée aux maladies cardiovasculaires, plusieurs programmes nationaux ont été mis en place afin de réduire les facteurs de risque ou augmenter les facteurs de protection dans la population en générale. Ils incluent : (i) Programme National de Réduction du Tabagisme (PNRT [1]) ; (ii) le Programme National Nutrition Santé (PNNS [2]) ; (iii) le Plan Santé Sport Bien-Etre (SSBE [3]) ; (iv) le Plan National Santé Environnement (PNSE [4]) ; (v) le Plan de santé au travail 2016-2020 [5].

Certains facteurs de risque tels que l'âge et les facteurs génétiques de maladies cardiovasculaires ne peuvent être modifiés. Cependant, ils doivent être pris en compte dans l'évaluation du risque. Par exemple, pour les femmes, l'âge devient un facteur de risque à 55 ans. Si votre père ou votre frère a eu une crise cardiaque avant l'âge de 55

1. https://solidarites-sante.gouv.fr/IMG/pdf/PNRT2014-2019.pdf
2. https://solidarites-sante.gouv.fr/IMG/pdf/PNNS_2011-2015.pdf
3. http://www.sports.gouv.fr/IMG/pdf/1-texte_communication_en_conseil_des_ministres.pdf
4. https://solidarites-sante.gouv.fr/IMG/pdf/2014-06-24-rapport_cas-vf.pdf
5. https://travail-emploi.gouv.fr/IMG/pdf/pst3.pdf

Catégories	Nom du facteur de risque	Effets sur les maladies cardiovasculaires
Facteurs de risque non modifiables	Age	Avec l'augmentation de l'âge, les vaisseaux sanguins ont tendance à perdre de l'élasticité et à accumuler les dépôts d'athérome. Ceci augmente le risque de faire une maladie cardiovasculaire.
	Antécédents familiaux/génétique	Les antécédents familiaux sont considérés comme un moyen indirect pour identifier les marqueurs génétiques augmentant le risque de faire une maladie cardiovasculaire. Si un parent proche, à développer une maladie cardiovasculaire précoce, vous avez un risque plus important d'en faire autant.
	Sexe	Les hormones sexuelles féminines sont un facteur protecteur contre la survenue des maladies cardiovasculaires. Avant la ménopause, les hommes sont plus à risque que les femmes.
Facteurs de risque modifiables	Cholestérol sanguin : HDL et LDL	Le cholestérol est vital pour les cellules saines. Lorsque le régime alimentaire est riche en graisses saturées, le corps à tendance à accumuler le cholestérol. Si cela ce produit une partie du cholestérol peut se déposer sur les parois des artères, créant ainsi des plaques d'athéromes.
	Triglycérides	La présence excessive de triglycérides dans le sang, associée à une quantité insuffisante de HDL (lipoprotéines de haute densité), est un facteur de risque cardiovasculaire. Cette combinaison est souvent liée à la survenue d'une athérosclérose, associée à un risque accru d'AVC et de cardiopathies ischémiques. L'hypertriglycéridémie peut être héréditaire mais survient également chez les personnes en surpoids, en particulier lorsque l'excès de graisse est localisé autour de la taille.
	Hypertension artérielle	L'hypertension artérielle augmente le risque de crise cardiaque, d'insuffisance cardiaque et d'accident vasculaire cérébral. La maladie survient plus fréquemment lorsqu'on a des antécédents familiaux d'hypertension artérielle, ou un mode de vie qui y prédispose (sédentarité, alimentation malsaine, etc.)
	Diabète sucré	Les personnes atteintes de diabète courent un risque beaucoup plus élevé de maladie cardiovasculaire. Ce risque accru est associé à une glycémie élevée, à une pression artérielle élevée et à une augmentation des lipides sanguins.
	Tabagisme	Le fait de fumer, même quelques cigarettes par jour, augmente le risque de maladie cardiovasculaire. Arrêter de fumer, même après avoir fumé pendant de nombreuses années, réduira le risque de MCV.

TABLE 1.1 – Liste des facteurs de risque cardiovasculaire

Facteurs de risque modifiables	Surpoids et Obésité	Le surpoids augmente le risque de maladie cardiovasculaire. Cela tient en partie au fait que les personnes en surpoids sont plus susceptibles d'avoir de l'hypertension artérielle, le diabète sucré et un trouble du métabolisme des lipides. Une perte de poids modeste peut réduire ces risques, et le bénéfice perdure tant que la perte de poids est maintenue.
	Sédentarité (inactivité physique)	L'inactivité physique est un contributeur important à la maladie cardiovasculaire. Les avantages cardiovasculaires d'une activité physique régulière incluent une pression artérielle réduite, un contrôle du poids, un tour de taille réduit qui contribuent tous à réduire le risque de développer une maladie cardiovasculaire.
	Consommation d'alcool excessif	Une consommation excessive d'alcool peut nuire à la santé de plusieurs façons. En effet, l'alcool, s'il est pris en excès, est nocif pour le cœur et d'autres organes. Il peut endommager directement le muscle cardiaque et provoquer des battements irréguliers du cœur. L'alcool contribue également à la prise de poids, à l'hypertriglycéridémie, à l'hypertension artérielle, et aux accidents vasculaires cérébraux. Il y a de nombreuses raisons de maintenir la consommation d'alcool dans des limites raisonnables, ou au mieux d'arrêter de consommer de l'alcool.
	Dépression	Les personnes dépressives sont plus susceptibles de développer une maladie cardiovasculaire. Il existe une relation dose-dépendante : plus la dépression est sévère, plus le risque d'événements cardiovasculaires est élevé.
	Troubles du sommeil	Les études épidémiologiques suggèrent la responsabilité de la réduction du temps de sommeil dans l'apparition de l'obésité et du diabète sucré. Elles montrent son association à un mauvais équilibre du diabète, à l'HTA, aux atteintes coronariennes, à l'inflammation et au syndrome métabolique. Les mêmes associations sont retrouvées avec l'insomnie ou la qualité du sommeil, indépendamment de la présence des principales comorbidités et des troubles respiratoires du sommeil.
	Alimentation malsaine	Le rôle de l'alimentation est crucial dans le développement et la prévention des maladies cardiovasculaires. L'alimentation est l'un des principaux facteurs de risque modifiables afin de réduire l'incidence de la plupart des facteurs de risque cardiovasculaires. Une alimentation riche en graisses saturées, en viande, en sel entre autres augmente le risque de maladie cardiovasculaire. On estime qu'une alimentation malsaine cause environ 31% des maladies coronariennes et 11% des accidents vasculaires cérébraux dans le monde.

TABLE 1.2 – Suite de la liste des facteurs de risque cardiovasculaire

ans, votre mère ou votre soeur avant l'âge de 65 ans, vous présentez un risque accru de développer une maladie cardiovasculaire.

Il est possible d'évaluer un niveau de risque cardiovasculaire pour un individu. Ce niveau de risque est directement proportionnel au nombre de facteurs de risque présents chez cet individu : on parle de risque cardiovasculaire global. La stratégie pour diminuer le risque global consiste à agir sur ces facteurs de risque. En d'autres termes, il faut intervenir sur les facteurs de risque modifiables des maladies cardiovasculaires (Piepoli *et al.*, 2016).

Un des avantages de prendre en compte l'ensemble des facteurs de risque identifié est de pouvoir disposer d'une évaluation du risque adaptée. Enfin, avant de mettre en place une approche d'évaluation du risque, il est important de comprendre la notion du risque et les approches existantes pour son évaluation et sa prévention. Cette analyse constitue une très bonne base à la compréhension des connaissances qui peuvent être intégrées dans un système d'évaluation du risque.

1.3 Prévention du risque

1.3.1 Définition du risque

Le risque est un sujet important dans la société. Les individus sont confrontés aux risques des marchés financiers, des centrales nucléaires, des catastrophes naturelles et des fuites dans les systèmes de Technologies de l'Information et de la Communication (TIC), pour ne citer que quelques exemples de domaines dans lesquels l'incertitude et le risque de préjudice jouent un rôle important. Il n'est donc pas surprenant que le risque soit étudié dans des domaines aussi divers que les mathématiques et les sciences naturelles, mais aussi la psychologie, l'économie, la sociologie, les études culturelles et la philosophie (Roeser *et al.*, 2012).

Selon (MÖLLER *et al.*, 2006 ; Roeser *et al.*, 2012), il est utile de distinguer au moins cinq significations du risque différentes, mais clairement liées, qui ont été utilisées dans la littérature :

1. le risque est un événement indésirable pouvant ou non se produire ;
2. le risque est la cause d'un événement indésirable pouvant ou non se produire ;
3. le risque est la probabilité d'un événement indésirable qui peut ou non se produire ;
4. le risque est le fait qu'une décision soit prise dans des conditions probabilistes ;
5. le risque correspond à la valeur statistique attendue des événements indésirables qui peuvent ou non se produire.

Dans le domaine de la santé, le risque est la probabilité de développer un événement (par exemple cardiovasculaire) précis chez des sujets donnés sur une période bien déterminée. La notion de risque est donc une notion populationnelle que l'on essaie de transposer à des individus. Or, toutes les évaluations du risque que nous possédons sont développées à partir d'échantillons issus de populations. Par conséquent, le risque communiqué à un individu reflète l'expérience d'un groupe et il est impératif de connaître l'origine du calcul. Il s'agit généralement d'une incidence, c'est-à-dire d'un taux de nouveaux événements cardiovasculaires chez des sujets auparavant sains. Il peut s'agir alors d'une incidence de décès, d'une incidence de sujets avec infarctus du myocarde avec ou sans décès, de sujets avec infarctus du myocarde ou de maladie coronaire au

sens large ou encore de sujets ayant présenté un événement cardiovasculaire, quel qu'il soit (Ferrières, 2012).

1.3.2 Définition de la prévention

La prévention peut être décrite comme un ensemble complexe d'activités visant à : (i) développer les compétences personnelles et sociales des personnes ; (ii) modifier les systèmes sociaux pour mieux répondre aux besoins des personnes. Travaillant à ces deux niveaux, la prévention englobe les actions visant les influences individuelles, sociales, économiques et culturelles, supposées contribuées à l'émergence de comportements dysfonctionnels individuels ou inversement et pouvant être modifiés pour contribuer positivement à l'émergence de comportement social (Adam, 1981).

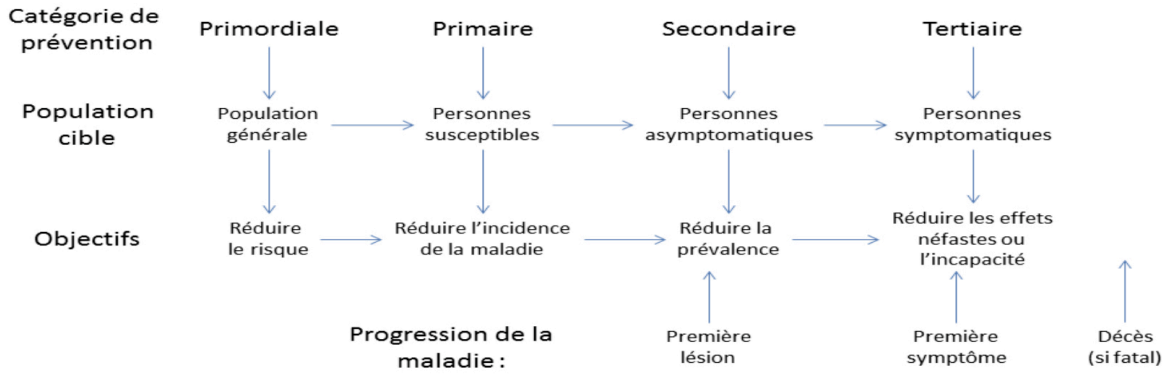


Figure 1.2 – Classification de stratégies préventives

Il existe différents types de prévention (voir figure 1.2) : (i) la première est la « prévention primordiale » qui consiste en des actions et des mesures qui empêchent l'émergence de facteurs de risque sous forme de conditions environnementales, économiques, sociales et comportementales et de modes de vie culturels, etc. ; (ii) la seconde est la « prévention primaire » qui désigne les mesures d'intervention visant à prévenir l'apparition primordiale d'une nouvelle maladie, d'un handicap ou d'une blessure ; (iii) la troisième est la « prévention secondaire » qui est un ensemble de mesures utilisées pour la détection précoce et une intervention rapide afin de contrôler un problème ou une maladie et d'en minimiser les conséquences ; (iv) la quatrième est la « prévention tertiaire » qui se concentre sur la réduction des complications ultérieures d'une maladie, d'un handicap ou d'une blessure existants, par le traitement et la réadaptation (Baumann et Ylinen, 2017). Dans la suite de ce document, la prévention traitée est la prévention primaire selon le principe « *mieux vaut prévenir que guérir* ». La principale raison de préférer la prévention primaire à la prévention secondaire est que tant que le danger existe toujours, il peut être réalisé par un événement déclencheur imprévu (Roeser *et al.*, 2012).

La prévention des maladies cardiovasculaires est définie comme un ensemble coordonné d'actions, au niveau de la population ou ciblant un individu, visant à éliminer ou à minimiser l'impact des maladies cardiovasculaires et de leurs handicaps associés (Feinleib, 2001). Selon la définition de l'OMS (WHO, 2007), la prévention concerne « toutes les activités dont le but principal est de promouvoir, restaurer et maintenir la santé ».

Les stratégies de prévention représentent des mesures préventives axées sur les personnes jugées les plus susceptibles de développer une certaine maladie. Elles sont pré-

cieuses dans la mesure de prévention au sein de la population générale. Un exemple d'une approche à haut risque pour la prévention des accidents vasculaires cérébraux est le traitement de l'hypertension chez les personnes ayant déjà subi un Accident Ischémique Transitoire (AIT). La stratégie à haut risque relève principalement du domaine médical de la prévention secondaire (Wilhelm, 2008).

1.3.3 Systèmes d'évaluation du risque

L'évaluation et la gestion des risques sont des approches abordées dans diverses disciplines, notamment la finance, le nucléaire, la santé, etc. La principale tâche de l'évaluation et de la gestion des risques consiste à mettre au point des outils adéquats pour faire face aux problèmes de complexité, d'incertitude et d'ambiguïté. Les systèmes informatisés d'évaluation des risques deviennent de plus en plus importants dans la société. Si les chances de réduire et de gérer certains risques augmentent, leurs impacts peuvent être contrôlés et réduits de manière significative (Andersson et Hossain, 2014).

1.3.3.1 Evaluation du risque dans d'autres domaines

Il existe de nombreux domaines dans lesquels l'évaluation du risque joue un rôle important. Nous avons selectionné les domaines dans lesquels la prévention du risque constitue un enjeu sociétal majeur.

Domaine des catastrophes naturelles. Le premier exemple est le domaine des catastrophes naturelles. Dans ce domaine, les inondations sont considérées comme l'une des formes les plus graves de catastrophes naturelles, responsables du plus grand nombre de décès et de dommages économiques. (Andersson et Hossain, 2014) proposent un modèle fondé sur des règles de croyance et diverses technologies de réseau de capteurs sans fil pouvant être utilisés en collaboration pour mettre en place des applications d'évaluation des risques.

Domaine du risque nucléaire. Le deuxième exemple est dans le domaine du risque nucléaire. ADAC (Capet *et al.*, 2008) est un système d'évaluation dynamique des risques qui suit l'évolution quotidienne de la situation dans un domaine particulier afin de permettre aux experts de mieux comprendre les situations à risque. Le système est fondé sur une bibliothèque de scénarios d'experts décrivant des développements de crise typiques et une ontologie représentant la connaissance du domaine. Il surveille un flux séquentiel d'événements entrants pour repérer les séquences d'événements susceptibles de produire une crise. Le système permet d'évaluer automatiquement les risques en matière de prolifération nucléaire. Il repose essentiellement sur deux composantes : l'analyse linguistique d'articles de presse et le calcul du risque nucléaire.

Domaine du risque chimique. Le troisième exemple est dans le domaine du risque chimique. (Nakai et Suzuki, 2014) proposent une méthode fondée sur le calcul de la probabilité d'occurrence des dangers dans les usines chimiques qui tiennent compte des garanties existantes. Le système développé repose sur l'analyse HAZOP (Hazard and Operability Study) et la fiabilité de l'agencement des équipements de sécurité. Le système vérifie si les garanties sont adéquates ou non et produit des recommandations pour une réduction supplémentaire des risques.

Domaine des transports. Le quatrième exemple est dans le domaine des transports.(Mazouni, 2008) propose une approche fondée sur la modélisation ontolo-

gique du processus accidentel pour mettre en place un système interactif d'aide à la décision. Il propose d'exploiter efficacement l'échange de savoir-faire en matière de gestion des risques relatifs à différents systèmes issus de différents domaines, et enfin de développer un Système Interactif d'Aide à la Décision (SIAD). Cette approche permet d'automatiser cette méthode et apporte des éléments d'aide à la décision en matière de gestion des risques à travers une interface graphique et ergonomique.

Domaine de la santé. Le cinquième exemple est dans le domaine de la santé. (Maserat *et al.*, 2017) proposent un système d'évaluation des risques fondé sur des preuves pour le dépistage du cancer. La précision du système de détection des groupes à risque est de 100%. Le système fondé sur des données probantes a estimé les taux de survie des populations couvertes avec une précision de 95,6%. (Atashi *et al.*, 2017) proposent un système d'inférence neuro-flou adaptatif (Adaptive Neuro-Fuzzy Inference System). Il s'agit d'un modèle informatique souple basé sur la précision des réseaux neuronaux et des avantages de prise de décision floue, ce qui peut grandement faciliter la modélisation diagnostique. L'approche a été utilisée pour la détection du cancer du sein.

Dans le domaine de la santé, les objets connectés sont de plus en plus utilisés. Ils permettent non seulement d'informer et d'alerter le personnel soignant, mais également d'identifier des problèmes de santé avant que ceux-ci ne deviennent critiques. L'objectif de ces derniers est d'améliorer diagnostics et traitements grâce aux données collectées, il devrait aussi contribuer à améliorer la prise en charge des patients grace à la visibilité de la progression d'une maladie. (Machorro-Cano *et al.*, 2017) proposent une approche intégrant différentes technologies de l'information et de la communication (TIC) et reposant sur des algorithmes d'intelligence artificielle permettant d'identifier des variables critiques fondées sur la surveillance des signes vitaux extraits de capteurs et de dispositifs permettant de prédire l'obésité, et l'hypertension et leurs complications possibles.

Le constat établi à partir de ces systèmes est que pour mettre en place un système d'évaluation du risque, il est nécessaire de : (i) modéliser les facteurs de risque du domaine considéré ; (ii) évaluer le risque par une approche globale en prenant en compte la totalité des facteurs et leurs interactions ; (iii) considérer l'ensemble des conceptualisations. Ces systèmes ne permettent cependant pas de déterminer la meilleure conceptualisation. Il est donc nécessaire de proposer une factorisation de ces conceptualisation afin de construire un modèle générique.

1.3.3.2 Evaluation du risque cardiovasculaire

L'évaluation du risque cardiovasculaire est une composante fondamentale de la prévention des maladies cardiovasculaires. Elle doit être réalisée avec une rapidité et une précision suffisantes. Les lignes directrices actuelles pour la prévention des maladies cardiovasculaires soulignent l'importance d'évaluer le risque cardiovasculaire global. Ce processus nécessite de prendre en compte les facteurs de risque modifiables tels que l'hypertension et les facteurs de risque non modifiables tels que l'âge et le sexe (De-Backer, 2003). Cette recommandation s'appuie sur des niveaux de preuves épidémiologiques élevés, qui montrent que les facteurs de risque cardiovasculaires ont des effets synergiques sur le niveau global de risque (Pletcher et Moran, 2017).

De nombreux systèmes ont été développés pour évaluer le risque cardiovasculaire d'un individu pour différentes populations. Le risque cardiovasculaire est évalué à l'aide

de différentes équations. La première est issue de l'étude épidémiologique de Framingham [6] développée vers 1967 par Cornfield et Truett (Truett *et al.*, 1967 ;DAgostino *et al.*, 2008). Depuis lors, l'équation a été réaménagée plusieurs fois, simplifiée par le SCORE (CONROY *et al.*, 2003), l'ASSIGN [7] (Woodward *et al.*, 2005), le Q-Risk (Hippisley-Cox *et al.*, 2007), le CUORE (Giampaoli, 2007), l'Arriba (Krones *et al.*, 2008) et le Globorisk (Hajifathalian *et al.*, 2015) pour ne citer que ceux là.

Cette diversité de système trouve son origine entre autres dans l'incapacité du score de Framingham à mesurer le risque cardiovasculaire chez les femmes plus jeunes. Cette situation a conduit les chercheurs à développer de nouveaux algorithmes d'évaluation du risque cardiovasculaire fondés à la fois sur les facteurs de risque traditionnels (ceux de l'équation de Framingham) et de nouveaux facteurs de risque (par exemple, l'obésité, les antécédents cardiovasculaires, etc.). Les calculs ont été effectués en utilisant des bases de données épidémiologiques construites dans ce but. Le choix d'un système d'évaluation du risque cardiovasculaire doit être fondé sur sa robustesse et sa capacité à inclure les facteurs de risque cliniquement pertinents, conduisant à un gain de santé mesurable (WHO, 2007).

Différentes études ont examiné la fiabilité de ces algorithmes utilisés pour évaluer le risque cardiovasculaire. (Torre *et al.*, 2018) proposent une étude qui examine la fiabilité de deux algorithmes utilisés pour évaluer le risque cardiovasculaire : l'algorithme de Framingham et l'algorithme CUORE [8]. Il est apparu que le score CUORE était moins fiable que le SCORE de Framingham ; en fait, dans le modèle de régression linéaire multiple, le coefficient de détermination était plus élevé lorsque la variable indépendante était l'échelle de Framingham pour le risque de coronaropathie.

Depuis 2003, les lignes directrices européennes sur la prévention des maladies cardiovasculaires en pratique clinique (Piepoli *et al.*, 2016) recommandent l'utilisation du système SCORE, car il a été développé à partir de vastes ensembles de données de cohorte européenne représentative (Aktas, 2004). En France, la HAS recommande d'évaluer le risque cardiovasculaire en prévention primaire à l'aide de l'outil SCORE (Systematic Coronary Risk Estimation). La version proposée évalue le risque de mortalité cardiovasculaire à 10 ans, en fonction du sexe, de l'âge (de 40 à 65 ans), du statut tabagique, de la pression artérielle systolique et des concentrations de cholestérol total (CT) contrairement à la version de base (SCORE).

Le système SCORE estime le risque à 10 ans d'un premier événement athéroscléreux mortel. Tous les codes de la Classification Internationale des Maladies (CIM) que l'on suppose être athérosclérotiques sont inclus, y compris la coronaropathie, les accidents vasculaires cérébraux et l'anévrisme. A l'aide de ce SCORE, les patients peuvent être classés en groupes à risque faible, modéré ou élevé.

Cependant, d'autres facteurs de risque sont bien connus : la sédentarité, l'alimentation, l'obésité, le stress par exemple. La plupart de ces facteurs de risque reposent sur des comportements, qu'il est bien difficile de mettre en équation. C'est pour cette raison que de nombreuses approches tentent d'évaluer de façon globale le risque cardiovasculaire en prenant en compte ces modificateurs du risque.

Le système clinique hybride basé sur une ontologie et l'apprentissage par machine développé par (Hussain, Farooq *et al.*, 2015) comporte deux composants clés : (i) un

6. Les algorithmes de risque actuels de Framingham comprennent l'âge, le sexe, le tabagisme, la pression artérielle et le taux de cholestérol sanguin (DAgostino *et al.*, 2008)

7. Cardiovascular risk estimation model from the Scottish Intercollegiate Guidelines Network

8. en insérant dans le questionnaire proposé par Giampaoli *et al.*, 2012, les variables sexe, âge, tabagisme, pression artérielle systolique, cholestérol total et HDL, présence de diabète et prise régulière d'antihypertenseurs, ont permis de calculer le score individuel.

système d'évaluation des risques cliniques et de recommandation fondée sur l'ontologie ; (ii) un système prédictif axé sur l'apprentissage machine. L'objectif principal de cette approche est de combiner les connaissances collectées *via* le système d'évaluation et les preuves issues du système de pronostic par apprentissage automatique pour offrir un cadre d'aide à la décision clinique global dans le domaine cardiovasculaire.

Parmi les techniques d'apprentissage automatique utilisées pour prédire les maladies cardiovasculaires, les réseaux neuronaux (NN) sont couramment utilisés pour améliorer la précision des performances de prédiction. (Narain *et al.*, 2016) proposent une approche prédisant le niveau de risque cardiovasculaire utilisant les bases de données sur les maladies cardiaques représentées sous forme d'un réseau de neurones quantiques (QNN). L'objectif est de prendre des décisions cliniques intelligentes qui ne peuvent pas être réalisées dans des systèmes d'évaluation du risque traditionnels (Framingham). (J. K. Kim et Kang, 2017) utilisent une analyse de corrélation des caractéristiques (NN-FCA) en deux étapes. La première étape est celle de la sélection des caractéristiques. Les caractéristiques les plus importantes pour la prédiction du risque de maladie coronarienne sont alors classées. La seconde étape est celle de l'analyse de la corrélation de ces caractéristiques. Les travaux de (W. *et al.*, 2017) ont montré que l'apprentissage automatique améliore considérablement la précision de la prévision du risque cardiovasculaire, en augmentant le nombre de patients identifiés pouvant bénéficier d'un traitement préventif (sensibilité), tout en évitant de traiter inutilement les autres (spécificité).

Plusieurs facteurs doivent être pris en compte pour évaluer le risque cardiovasculaire, comme le montre le tableau 1.1. La décision sur la manière de procéder au cours de l'évaluation dépend du patient. Clairement, toutes les méthodes ne doivent pas être utilisées pour établir le risque individuel (Feitosa, 2015).

Le sexe, l'âge et les antécédents familiaux de maladies cardiovasculaires ne peuvent être modifiés, mais il convient de tenir compte de ces facteurs lorsque des actions préventives sont discutées. En utilisant par exemple le tableau de risque SCORE (CONROY *et al.*, 2003), il est facile de voir que l'âge joue un rôle important dans la prédiction du risque : un homme de 40 ans ne présente qu'un risque modéré, mais avec les mêmes niveaux de risque, il pourrait bien être une personne à risque élevé à 60 ans (Wilhelmsen, 2017).

Cependant, l'évaluation du risque cardiovasculaire demeure difficile. Plusieurs questions peuvent justifier les échecs des approches d'évaluation du risque cardiovasculaire : (i) les cliniciens utilisent-ils les bons facteurs de risque cardiovasculaire et les bonnes équations et outils d'évaluation des risques ? ; (ii) comment faciliter l'évaluation du risque cardiovasculaire pour les décideurs cliniques occupés, débordés d'informations et comment les aider à communiquer les niveaux de risques aux patients ? ; (iii) les cliniciens peuvent-ils utiliser la technologie et d'autres progrès biologiques pour améliorer l'évaluation du risque cardiovasculaire et cibler avec précision la prévention pour les patients à haut risque ?

D'un point de vue applicatif, il existe des preuves que, malgré l'acceptation généralisée du concept de risque cardiovasculaire global, l'évaluation du risque n'est pas effectuée systématiquement dans la pratique clinique. De plus, les seuils pour déclencher certaines interventions sont problématiques, car le risque est un *continuum* et il n'y a pas de seuil auquel, par exemple, un médicament est automatiquement recommandé. De toute évidence, les décisions concernant l'initiation du traitement doivent également être basées sur les préférences du patient. Un problème particulier concerne les jeunes présentant des niveaux élevés de facteurs de risque et un score de risque faible

pouvant masquer un risque relatif très élevé à long terme. Cette situation nécessite des conseils intensifs en matière de style de vie.

1.3.4 Approches de prévention sensibles au domaine

Une fois l'évaluation des risques terminée, l'étape suivante consiste à formuler des plans de prévention des risques efficaces, axés sur la probabilité d'un résultat donné ou de plusieurs résultats. Le plan de prévention des risques doit indiquer clairement et précisément la nature et le niveau du risque. Dans le but de minimiser les risques, une attention accrue a été accordée au développement d'outils d'évaluation appropriés, valides et fiables et à la création de stratégies nouvelles et plus efficaces pour gérer les défis qui découlent inévitablement d'une prise de risque accrue.

Dans le cas de la pollution par exemple, la prévention consiste à réduire ou à éliminer la toxicité des déchets. La loi sur la prévention de la pollution de 1990 établit une hiérarchie des approches privilégiées de protection de l'environnement. En premier lieu, la pollution devrait être évitée à la source chaque fois que cela est possible. Si les flux de déchets ne peuvent pas être évités, ils doivent être réutilisés, recyclés ou traités. L'élimination doit être le dernier recours (Wilhelm, 2008).

Dans le cas des maladies cardiovasculaires, l'approche consiste à cibler chacun des facteurs de risque identifiés lors de l'évaluation par un ensemble de recommandations. D'une part, les lignes directrices (OMS, HAS, PNNS, etc.) sont une source de référence pour réduire le profil de risque de maladies cardiovasculaires avec un ensemble de recommandations. D'autre part, la plupart de ces facteurs de risque reposent sur des comportements, qu'il est bien difficile de mettre en équation. L'objectif est de trouver un compromis entre la motivation à des changements de comportement et son niveau de risque. De plus, le plan de gestion des risques n'est qu'une stratégie à court terme et doit être revu constamment et méthodiquement par toutes les personnes responsables et impliquées dans la prise en charge des utilisateurs du service.

Malgré les connaissances disponibles sur l'impact des facteurs de risque cardiovasculaires et sur l'existence de stratégies de prévention primaire efficaces, l'adoption d'habitudes saines par la population est encore très faible. Le projet de prévention cardiovasculaire de la Carélie du Nord (Vartiainen, 2018) est né du besoin des Finlandais de réduire la forte mortalité due aux maladies cardiovasculaires. En 40 ans, la mortalité a diminué de 80% grâce à l'intervention combinée des changements environnementaux et de la réorganisation des services de santé, la formation des professionnels et l'utilisation des médias. Cela montre que la prévention cardiovasculaire est une question de travail en équipe, visant tous les facteurs de risque et nécessitant une collaboration avec des spécialistes de différents domaines.

Finalement, la prévention doit s'intéresser à tous les facteurs pouvant être mis en cause dans la genèse d'un risque ou d'une maladie : il s'agit de les analyser pour déceler l'importance de leurs effets isolés et conjugués, et de trouver des mesures et des moyens pour les éradiquer si possible, sinon les rendre moins influents. Les approches sont différentes selon le domaine et selon les objectifs à atteindre. Ainsi, nous pouvons avoir des approches de prévention à plusieurs couches comme c'est le cas dans le risque nucléaire, car la conséquence d'un risque peut être dramatique.

1.3.5 Vers un système de prévention du risque cardiovasculaire informatisé

Les modifications du mode de vie dans le but de réduire le risque cardiovasculaire consistent principalement à lutter contre le tabagisme et l'obésité, à pratiquer une activité physique régulière et à avoir une alimentation saine. Ils ont un impact significatif sur la prévention primaire des maladies cardiovasculaires. L'indicateur de l'espérance de vie ajustée sur sa qualité (QALY) est un outil important pour la quantification de l'impact d'une maladie ou d'une intervention, car il combine la quantité et la qualité de vie en un seul score. Malgré les connaissances disponibles sur l'impact des facteurs de risque cardiovasculaires et sur l'existence de stratégies de prévention primaire efficaces, l'adoption d'habitudes saines par la population est encore très faible.

Pour tenter de répondre aux problématiques exposées dans l'introduction, nous avons choisi d'adopter une approche hybride qui exploitera à la fois les connaissances concernant les facteurs de risque et les techniques issues du web sémantique. La représentation des connaissances et la puissance du raisonnement offert par les standards du web sémantique seront au cœur de notre méthode. De cette manière, nous serons capables de réaliser une évaluation du risque cardiovasculaire précise qui permettra une prise en charge flexible des interactions qui existent entre les facteurs de risque cardiovasculaire.

De manière détaillée, nous proposons d'extraire et de représenter les connaissances sur les interactions entre les facteurs de risque cardiovasculaire sous forme d'un modèle conceptuel à partir de données statistiques présentées sous forme de tableaux. Nous avons extrait les connaissances sur la prise en charge et les recommandations à partir des guides de bonne pratique. L'application du raisonnement sur ces connaissances permettra de construire et d'améliorer la précision et la pertinence des évaluations et des recommandations. La soumission d'un profil patient ou individu à une telle base permettra l'obtention de recommandations adaptées qui seront présentées de manière claire et synthétique. En effet, quand il s'agit de systèmes de prévention, la seule pertinence des recommandations ne suffit pas à convaincre et l'ergonomie d'un système est souvent déterminante. Pour atteindre ce résultat, nous avons opté pour une approche de visualisation adaptée aux interactions homme-machine (voir chapitre 4).

1.4 Le web sémantique

Parler du web sémantique c'est avant tout parler du web. C'est pourquoi, il nous paraît intéressant de présenter un bref historique de ce dernier afin de comprendre le contexte dans lequel il est né (section 1.3.1) et de donner quelques notions essentielles (section 1.3.2) pour la compréhension du contexte des travaux présentés dans cette thèse.

1.4.1 Présentation générale du web sémantique

Le web sémantique est défini par (Hendler et Lassila, 2001) comme « une extension du web courant, dans laquelle on attribue à une information un sens bien défini pour permettre aux ordinateurs et aux gens de travailler en coopération » [9]. En d'autres termes, il s'agit d'identifier chaque donnée, de l'enrichir avec des métadonnées et de

9. « an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation »

la relier à d'autres données sur le web. En proposant cette vision d'interopérabilité [10] basée sur les standards du web et sur des liens entre les ressources, plusieurs systèmes peuvent communiquer entre eux, échanger et partager l'accès à l'information.

L'expression « web sémantique » est composée de deux mots : web et sémantique. Le web est défini comme étant un système hypertexte [11] public permettant de consulter des pages accessibles sur le réseau Internet. Le mot web est très important pour comprendre les enjeux du web sémantique. Depuis son lancement aux débuts des années 90, le web n'a cessé d'évoluer (voir figure 1.3).

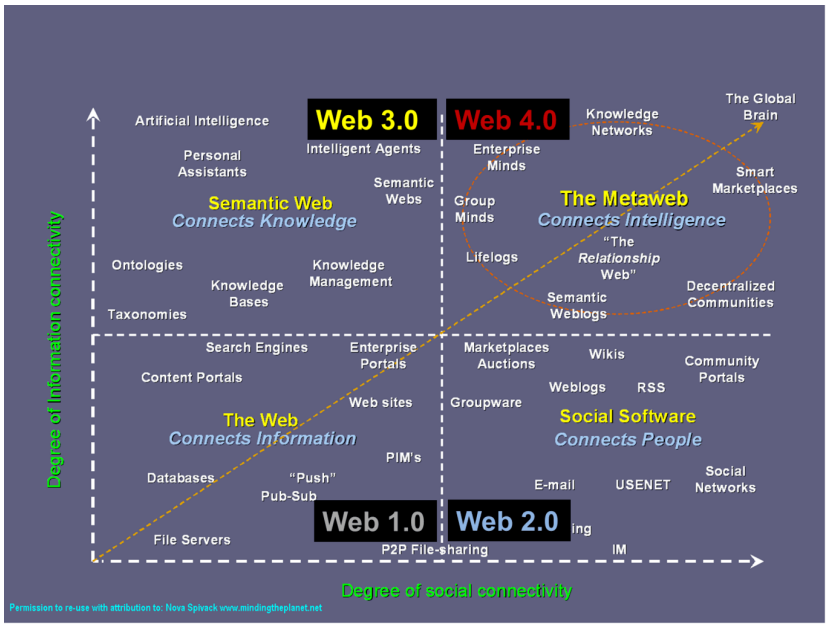


FIGURE 1.3 – Evolution du web (source : `https://c-marketing.eu/du-web-1-0-au-web-4-0/`)

— La première version du web est celle du « web 1.0 » qui peut se situer au début des années 90. Selon (BACHIMONT *et al.*, 2011), le web 1.0 « est le web dit documentaire, où les ressources sont publiées sans autre traitement que leur mise en forme et sans autre interactivité que l'activation des liens hypertextuels. ». Il se présente sous forme d'une production de contenus permettant aux entreprises de diffuser leurs informations aux utilisateurs. Les premiers sites web fournissaient un nombre limité d'interactions utilisateur ou de contributions au contenu et ne permettaient que de rechercher l'information et de la lire.

— La seconde version est celle du « web 2.0 », qui va être officiellement adoptée en 2004. Cette version change totalement de perspective et privilégie la dimension de partage et d'échange d'informations et de contenus (textes, vidéos, images ou autres). Au sein du web 2.0, nommé plus justement « web social » ou « web

10. « La capacité pour deux (ou plus) systèmes ou composants à échanger des informations et à utiliser les informations échangées (IEEE, 1991) »

11. Le concept d'hypertexte a été inventé à l'origine par Ted Nelson (NELSON, 1965), mais cette vision a été précedement utilisée en 1945 lorsque Vannevar Bush conseiller du président FD Roosevelt, a proposé la conception d'un système appelé MEMEX pour faciliter l'accès à de grandes quantités de données de manière associative. Dans son article, il a décrit les principes de construction de sa machine Memex à l'aide de microfilms, de relais électriques, de capteurs optiques et d'autres outils nécessaires pour accéder à une vaste collection de données grâce à des références croisées (WINSTEAD et GERSHON, 1996).

collaboratif », les usagers entrent dans un mode actif. Au fur et à mesure de leur navigation, les utilisateurs peuvent ajouter du contenu à travers des liens hypertextes et d'autres tags, annotations ou commentaires. Le contenu est produit par les utilisateurs grâce à l'émergence des blogs, des wikis, etc. L'utilisateur devient alors source d'informations et de création.

— La troisième version est celle du « web 3.0 » ou « web sémantique », qui désigne le web qui suit le web 2.0. Il vise à organiser la masse d'information disponibles en fonction du contexte et des besoins de chaque utilisateur, en tenant compte de sa localisation, de ses préférences, etc. Cependant, le web sémantique a été pensé en 1994 soit avant le web 2.0 (Berners-Lee, 1994), mais il a fallu attendre quelques années pour que les outils et les standards du web sémantique soient développés. Nous pouvons dire que le web sémantique et le web 2.0 ont évolué parallèlement depuis la fin des années 90.

— La quatrième version est celle du « web 4.0 » évoqué par certains comme le « web intelligent ». L'objectif est de permettre un accès au site web simultané en lecture-écriture-exécution-relationnel avec des interactions intelligentes, mais il n'existe pas de définition précise. Le web 4.0 est également connu sous le nom de web symbiotique dans lequel un esprit humain et des machines peuvent interagir en symbiose.

La « sémantique » est généralement synonyme de « sens des mots, des phrases ou des symboles ». En linguistique, la sémantique représente la science du sens sous différentes formes de langage. En informatique, et en particulier dans le domaine du web sémantique, la sémantique signifie la signification de mots ou de signes (chaînes) et leurs relations mutuelles. L'informatique, depuis ses débuts, s'est intéressée au traitement des données. Les langages de programmation fournissent des types de données simples et complexes pour stocker des données. A l'origine, la sémantique de ces données était câblée dans les programmes dans lesquels elles étaient interprétées et utilisées. Il y a environ 50 ans, les données ont commencé à être séparées du programme d'application pour être stockées dans des bases de données. Cela a permis de réutiliser les mêmes données dans différents contextes de programmation et d'empêcher la réimplémentation du même composant de gestion de données dans de nombreuses applications. Le fait que la signification des données ne soit plus câblée directement dans le programme d'application a conduit à développer des mécanismes de représentation de la structure et de la sémantique des données en cours de développement.

Le web sémantique est fondé sur des technologies incluant des langages de représentation des connaissances pour les ontologies ainsi que des méthodes et des outils pour créer, maintenir et appliquer des raisonnements sur ces ontologies. Les applications de ces technologies ne sont plus limitées au web, mais couvrent de nombreux domaines de l'informatique tels que la gestion des connaissances, l'informatique, l'intégration de données, les systèmes cognitifs, l'ingénierie logicielle, etc. A l'heure actuelle, il semble probable que les applications commerciales des technologies sémantiques concerneront d'abord des domaines qui ne font pas partie du web, mais représentent plutôt des solutions industrielles pour la gestion des connaissances.

L'objectif du web sémantique est de répondre à des besoins et exigences en se fondant sur le principe que certaines classes de tâches peuvent être déléguées à des machines et par conséquent traitées automatiquement. Citons à titre d'exemple, un chercheur qui présente ses travaux de recherche dans une conférence. Pour obtenir ses billets, il doit d'abord se renseigner sur les modes de transport disponibles à partir

de son lieu de départ et de son lieu de destination. Il doit ensuite consulter différents sites de vente de compagnies de transport pour pouvoir obtenir et comparer les différents tarifs, les horaires de déplacement et les préférences de confort. Ces tâches répétitives peuvent être traitées efficacement par des agents logiciels. Toutefois, pour que ces agents puissent parvenir à les accomplir, il est nécessaire que les données utiles soient disponibles en ligne dans un format compréhensible par une machine. C'est précisément dans ce but que le Web Consortium (W3C [12]) a déjà adopté des normes de base pour les langages de spécification d'informations, et qui seront présentés dans les chapitres suivants.

1.4.2 Les standards du W3C

En 1994 est créé le World Wide Web Consortium, ou W3C, qui va jouer un rôle primordial dans la normalisation du web. L'activité web du CERN (appelée projet WebCore) est alors transférée à l'INRIA, qui devient membre fondateur du W3C avec le MIT (*Massachusetts Institute of Technology*) aux États-Unis et l'Université de Keio au Japan (Gandon *et al.*, 2012). L'objectif du W3C est de proposer des standards pour la représentation, la visualisation et la diffusion des connaissances ainsi que pour le raisonnement.

Les standards proposés par le W3C sont organisés en couches pour construire un web de données liées et un web sémantique. La figure 1.4 présente cette architecture appelée *Semantic Web Layer Cake* ou pile des standards du web sémantique, que nous allons présenter brièvement. Cette disposition sous forme couche permet d'exploiter les possibilités offertes par les couches inférieures. Nous conseillons au lecteur de consulter les références citées dans ce chapitre pour obtenir plus de détails.

— **Identifier :** son but est de fournir un nom unique, qui fait référence à chaque ressource sur le web. Il repose sur les URI (*Universal Resource Identifier*), si de plus cet identifiant donne un chemin d'accès pour obtenir une représentation de cette ressource, alors on parle d'URL (*Universal Resource Locator*), connue sous le nom d'adresse web. Un autre format d'identifiant est celui d'IRI, il s'agit d'URI internationalisée, c'est-à-dire des adresses autorisant l'utilisation de divers alphabets de différentes langues du monde en utilisant le codage UTF-8 (Gandon *et al.*, 2012).

— **Syntax :** elle permet de décrire des ressources, favorisant ainsi les échanges sur le web. Plusieurs syntaxes existent, XML la plus populaire, permet de décrire les ressources à l'aide de balises (par exemple, auteur, date et acteurs d'un film).

— **Data interchange :** cette couche est la première brique des standards du web sémantique et recouvre à la fois un modèle et plusieurs syntaxes. Elle réutilise le mécanisme des URI pour identifier les ressources décrites et représenter explicitement toutes relations entre deux ressources. Elle repose sur le *framework* RDF (Resource Description Framework), que nous allons présenté dans la section (1.4.2.1).

— **Querying :** la recommandation SPARQL (présenté plus loin dans la section 1.4.2.3) est un langage destiné à interroger les connaissances représentées par le *framework* RDF. Ainsi, la combinaison de RDF et SPARQL peut être considérée

12. Le Web Consortium (W3C) est une communauté internationale où les membres, une équipe à plein temps, et le public travaillent ensemble pour développer les standards du web.

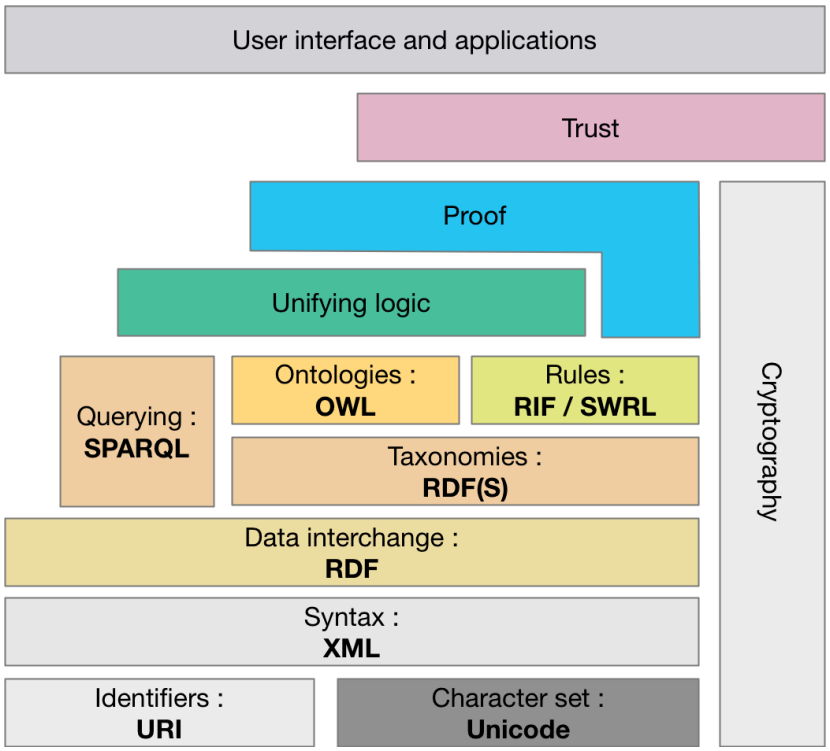


FIGURE 1.4 – Semantic Web Layer Cake ou la pile des standards du web sémantique (GANDON *et al.*, 2012)

comme la première phase de déploiement massif du web sémantique (GANDON *et al.*, 2012).

— **Ontologies :** Pour s'assurer de l'interprétation et de l'utilisation qui peut être faite des données, il est aussi utile de publier les vocabulaires ou les schémas (ou ontologie) qui lui sont associées (GANDON *et al.*, 2012). Un vocabulaire est une liste de termes contrôlés organisés de façon hiérarchique permettant de créer des classifications dans un domaine. Selon la définition de (GRUBER, 1995), une ontologie est « une spécification explicite d'une conceptualisation ». L'ontologie contient une description formelle des concepts d'un domaine, ainsi que les relations entre ceux-ci (MEILENDER, 2013). RDFS et OWL (présentés respectivement dans les sections 1.4.2.1 et 1.4.2.2) sont les deux briques suivantes dans les standards du web sémantique qui permettent de décrire les connaissances d'un domaine (voir figure 1.4). Plus le niveau d'abstraction est elevé plus l'expressivité des connaissances croît et plus le coût (en terme de matériel) du raisonnement sur ces schémas est elevé.

— **Rules :** son but est de permettre de représenter et d'échanger des connaissances dont l'expression n'est pas possible en logiques de descriptions sur lesquelles repose OWL. Ces règles logiques sont sous forme de « SI une hypothèse ALORS une conclusion peut être déduite ».

— **Unifying logic and Proof :** l'objectif de ces couches est fournir des outils nécessaires au raisonnement. Ainsi, un système de raisonnement utilisant les ontologies peut inférer de nouvelles connaissances.

— **Trust :** le but de cette couche est de fournir à l'utilisateur la possibilité de tracer, de certifier et de vérifier la source des informations pour décider si oui ou non ces informations sont fiables.

— **User interface and applications :** pour construire des applications complètes, il est nécessaire d'intégrer d'autres standards du web à celui du web sémantique pour proposer de nouveaux service, interactions, etc.

1.4.2.1 La base de tout : RDF

RDF est la première pile des standards du web sémantique et recouvre à la fois un modèle et plusieurs syntaxes pour représenter des données et des métadonnées à propos de tout sur le web (Gandon *et al.*, 2012).

Triplet : la plus petite structure de description en RDF est le triplet (*sujet*, *prédicat*, *objet*). Le sujet est la ressource que l'on souhaite décrire, représentée par son identifiant unique, c'est-à-dire une URI. Le prédicat, également présenté sous forme d'URI, définit la propriété qui s'applique à la ressource. Il établit une relation entre le sujet et l'objet en faisant de l'objet une caractéristique du sujet. Quant à l'objet, il peut prendre deux formes : soit une ressource présentée sous forme d'URI, soit un littéral, c'est-à-dire un objet typé, par exemple une chaîne de caractères ou un nombre entier.

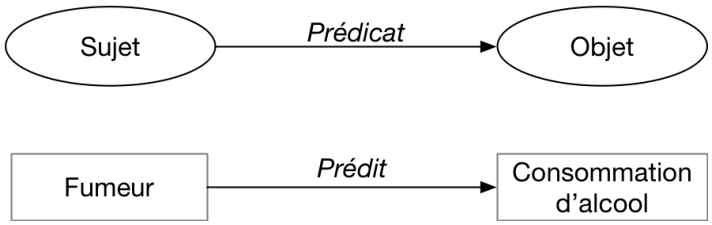


Figure 1.5 – Exemple de triplet RDF

Le point fort de RDF réside dans sa capacité à expliciter la nature de relation entre les deux ressources. Autrement dit, il s'agit non seulement de lier les données mais surtout d'interpréter le lien qui existe entre elles. Prenons à titre d'exemple, l'affirmation « fumeur prédit la consommation d'alcool », elle pourra être représentée par le triplet de la figure 1.5 où :

```
(http://www-limics.smbh.univ-paris13.fr/ontologie.owl#Fumeur,
http://www-limics.smbh.univ-paris13.fr/ontologie.owl#Predit,
http://www-limics.smbh.univ-paris13.fr/ontologie.owl#Consommation_Alcool)
```

où

— « http ://www-limics.smbh.univ-paris13.fr/ontologie.owl# » est l'espace de nom de l'ontologie des facteurs de risque cardiovasculaires ;

— « Fumeur » est la ressource représentant le facteur de risque fumeur qui correspond au sujet du triplet ;

— « Predit » est le prédicat signifiant « prédit » ;

— « Consommation_Alcool » est la ressource représentant le facteur de risque consommation d'alcool qui correspond à l'objet du triplet.

Syntaxe : Pour pouvoir représenter le modèle des graphes RDF en machine, il faut respecter une syntaxe concrète. Il existe plusieurs notations pour les graphes RDF, comme l'illustre la figure 1.6. Ces trois syntaxes sont des standards validés par le W3C : la syntaxe RDF/XML (Klyne et Carrol, 2004), la plus commune,

proposée dans le but de sérialiser les documents RDF en XML (voir figure 1.6(a)), la syntaxe N3 (Berners-Lee et Connolly, 2008), conçue pour favoriser la lisibilité par les humains (voir figure 1.6(b)) et la syntaxe JSON-LD (Sporny *et al.*, 2018), conçue pour s'intégrer facilement dans les systèmes déployés utilisant déjà le format JSON (voir figure 1.6(c)).

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:od="http://
www-limics.smbh.univ-paris13.fr/ontologie.owl#">
  <rdf:Description rdf:about="od:Fumeur">
      <od:Predit rdf:resource="od:Consommation_Alcool"/>
  </rdf:Description>
</rdf:RDF>
```

(a) Exemple de RDF avec la syntaxe RDF/XML

```
@prefix rdf : <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix od :  <http://www-limics.smbh.univ-paris13.fr/ontologie.owl#>.
<od: Fumeur>  od: Predit <od: Consommation_Alcool>.
```

(b) Exemple de RDF avec la syntaxe N3

```
{
  "@context": {
    "od": "http://www-limics.smbh.univ-paris13.fr/ontologie.owl#",
    "rdf": "http://www.w3.org/1999/02/22-rdf-syntax-ns#",
    "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
    "xsd": "http://www.w3.org/2001/XMLSchema#"
  },
  "@id": "od:Fumeur",
  "od:Predit": {
    "@id": "od:Consommation_Alcool"
  }
}
```

(c) Exemple de RDF avec la syntaxe JSON-LD

Figure 1.6 – Exemples de syntaxes RDF : la syntaxe RDF/XML, la syntaxe N3 et la syntaxe JSON-LD

Le vocabulaire RDFS : RDFS (RDF Schema) est le premier niveau des schémas légers, permettant de déclarer et de décrire les types de ressources manipulées (appelées classes : par exemple les facteurs de risque, les films, les recettes, etc.) et les types de relations entre ces ressources (appelées propriétés : par exemple, « a pour catégorie », « a pour titre », etc.). Ainsi, RDFS nous permet de définir des vocabulaires utilisés dans les graphes RDF et d'en nommer les primitives avec des URI : nommer les classes de ressources existantes ; nommer les types de relations existant entre les instances de ces classes et donner leurs signatures, à savoir le type de ressources qu'elles connectent ; organiser ces types dans des hiérarchies (Gandon *et al.*, 2012). Il permet d'utiliser ces signatures et les hiérarchies de classes et de relations pour automatiser des déductions. RDFS permet de déclarer le squelette hiérarchique d'un schéma dans un langage et avec des identifiants universels échangeables et réutilisables notamment pour permettre l'interopérabilité des systèmes.

1.4.2.2 Les langages de représentation OWL

La recommandation OWL (Ontology Web Language) a été développée par le W3C (D. L. McGuinness et Van-Harmelen, 2004). Elle est fondée sur RDFS et destinée

à la représentation des connaissances sous forme d'ontologies. Elle étend le vocabulaire RDFS en proposant des propriétés et des classes qui permettent d'exprimer des relations plus complexes. Le langage OWL est un langage de représentation des connaissances avec une grande expréssivité qui peut être utilisé pour modéliser des connaissances de base d'un domaine. Un axiome est une déclaration de base d'une ontologie OWL représentant un élément de connaissance. Par exemple, un énoncé comme « la classe facteur de risque modifiable est une sous-classe de la classe catégorie des facteurs de risque » est un axiome. Un axiome utilise des classes, des propriétés et des instances pour définir de nouvelles classes et propriétés plus complexes. Plusieurs syntaxes sont disponibles pour OWL et des outils permettant la traduction d'une syntaxe à une autre. RDF/XML est la syntaxe la plus courante et recommandée pour les documents OWL.

La recommandation OWL dans sa première version 1.0 (D. L. McGuinness et Van-Harmelen, 2004) définit trois fragments d'expressivité croissante, dont les deux premières couches sont fondées sur une famille de langage appelée « logique de description » permettant des raisonnements, vérification de la cohérence d'un schéma, la classification automatique des types afin de produire des hiérarchies, etc. Ces trois couches sont :

— OWL Lite est destiné aux utilisateurs ayant principalement besoin d'une hiérarchie de classes et de contraintes simples. Il inclut une forme limitée de restrictions de cardinalité. Les restrictions de cardinalité OWL (et OWL Lite) sont appelées restrictions locales, car elles sont indiquées dans les propriétés par rapport à une classe particulière. C'est-à-dire que les restrictions limitent la cardinalité de cette propriété aux instances de cette classe. Les restrictions de cardinalité d'OWL Lite sont limitées car elles ne permettent que des instructions concernant les cardinalités de valeur 0 ou 1 (elles ne permettent pas de valeurs arbitraires pour la cardinalité, comme c'est le cas dans OWL DL et OWL Full) (D. L. McGuinness et Van-Harmelen, 2004).

— OWL DL est destiné aux utilisateurs ayant besoin d'une expressivité maximale tout en gardant la complétude du calcul (toutes les inférences sont garanties calculables) et la décidabilité (tous les calculs s'achèveront dans un intervalle de temps fini) d'un point de vue théorique. En pratique, certaines ontologies représentées en OWL DL (par exemple GALEN [13]) cause des difficultés à certains raisonneurs qui n'arrivent pas à conclure ou donnent des résultats erronés (Alaya *et al.*, 2016). OWL DL comprend toutes les structures de langage OWL qui ne sont toutefois utilisables qu'avec certaines restrictions (par exemple, bien qu'une classe puisse être une sous-classe de plusieurs classes, elle ne peut pas être une instance d'une autre).

— OWL Full est destiné aux utilisateurs voulant une expressivité maximale et la liberté syntaxique de RDF sans garantie de calcul. Par exemple, dans OWL Full, une classe peut être simultanément envisagée comme une collection d'individus (classes) ou comme un individu à part entière. OWL Full permet à une ontologie d'augmenter la signification du vocabulaire prédéfini (RDF ou OWL). Un système de raisonnement ne pourra probablement pas gérer un raisonnement complet de toutes les caractéristiques de OWL Full.

L'un des principaux avantages de l'utilisation du langage OWL et d'autres langages de modélisation basés sur la logique est que ces langages prennent en charge de nombreux services de raisonnement. Ces dernières années, plusieurs moteurs d'inférence

13. https ://bioportal.bioontology.org/ontologies/GALEN

pour le raisonnement sur OWL ont été développés (Parsia *et al.*, 2017). Le principe consiste à traiter la connaissance (explicite) en arrière-plan et à déduire des informations implicites. Par exemple, considérons les déclarations, chaque père est un homme et Gilles est un père, alors un raisonneur OWL peut raisonner avec cette connaissance et en déduire que Gilles est un homme. La sémantique OWL soutient l'hypothèse du monde ouvert correspondant au principe selon lequel la connaissance du monde est incomplète, en d'autres termes, les déclarations qui ne sont pas connues pour être vraies ne sont pas nécessairement fausses.

La norme OWL 2 : la version actuelle de OWL appelée OWL 2 (D.-L. McGuinness et Van-Harmelen, 2012) proposée par le W3C en 2012 (Motik *et al.*, 2012), est plus expressive que sa version OWL 1. Elle offre une liste de nouvelles fonctionnalités, qui peuvent être classées dans les cinq grandes catégories suivantes :

1. améliore l'expressivité avec de nouvelles propriétés, telles que la propriété réflexive, la propriété irréflexive, etc ;
2. prise en charge étendue des types de données en permettant aux utilisateurs de définir leurs propres types de données lors de la création d'ontologies ;
3. étendre la capacité d'annotation. Plus précisément, il est possible d'ajouter des annotations aux axiomes, des informations de domaine et également des informations d'annotation aux annotations elles-mêmes ;
4. offre de nouveaux sous-langages : OWL 2 EL, OWL 2 QL et OWL 2 RL permettant un compromis entre expressivité et efficacité et offrant donc plus de choix aux utilisateurs.

1.4.2.3 Le langage d'interrogation SPARQL

SPARQL est un langage d'interrogation de graphe RDF conçu par le W3C (Prud'hommeaux et Seaborne, 2008). De manière schématique, nous pouvons dire que SPARQL est un langage de requête qui partage de nombreuses fonctionnalités avec d'autres langages de requête tels que XQUERY pour XML et SQL pour les bases de données relationnelles. L'énoncé de base de SPARQL est aussi un triplet (ressource, propriété, valeur). La syntaxe adoptée dans la version 1.0 de SPARQL est proche de celle de SQL et dispose des clauses SELECT et WHERE. Enfin, comme pour le langage SQL nous pouvons appliquer des traitements au résultat comme limiter le nombre de réponses ou trier selon un ou plusieurs critères. Par exemple, si nous cherchons à savoir quels sont les facteurs de risque qui prédisent le facteur « fumeur », la requête sera celle présentée dans la figure 1.7.

```
PREFIX od: <http://www-limics.smbh.univ-paris13.fr/ontologie.owl#>

SELECT?facteurs
WHERE
        {?facteurs od:predit od:Fumeur}
```

Figure 1.7 – Exemples de requête SPARQL

1.4.2.4 Le web de données

La notion du « web des données » a été introduite par (Berners-Lee, 2006) comme une nouvelle manière de publier les données structurées sur le web. Le web

sémantique ne consiste pas seulement à mettre des données sur le web. Il s'agit de créer des liens, de sorte qu'une personne ou une machine puisse explorer le réseau de données. Contrairement au web documentaire où les relations sont des ancres dans les documents hypertextes, dans le web des données les relations sont entre URIs. Au lieu d'avoir une page web pointant vers une autre page, un élément de données peut pointer vers une autre page en utilisant les URI.

L'exploitation du web en tant que plateforme d'intégration de données et d'informations ainsi que de recherche et d'interrogation est l'un des plus grands défis dans le domaine de la gestion intelligente des données. Le développement du web sémantique au cours des années 2000 a conduit à la conception d'une multitude d'ontologies, formant des ensembles de connaissances dans différents domaines. Cette évolution a conduit à l'apparition d'un web sémantique séparé du web classique, le premier étant destiné principalement aux échanges entre machines et le second aux échanges entre humains. Pour briser cette frontière, la première évolution proposée par le W3C est RDFa [14]. RDFa consiste à intégrer directement du contenu RDF dans des attributs HTML publié dans le web. Le code RDFa, est invisible pour l'utilisateur, mais indique et permet l'extraction automatique de données présentes dans une page web.

Le terme « données liées » fait ici référence à un ensemble de meilleures pratiques pour la publication et la connexion de données structurées sur le web. Ces dernières ont été adoptées par un nombre croissant de fournisseurs de données au cours des dernières années, conduisant à la création d'un espace global de données. Différents projets visant à publier et donner accès à des données ouvertes et liées ont vu le jour. Le projet Linking Open Data (LOD) lancé par le groupe SWEO [15] du W3C en 2007 est parmi ces projets. Ils visent à mettre à disposition sur le web des données au format RDF de diverses sources ouvertes et à ajouter des liens entre ces données. LOD a connu une croissance considérable au cours des dernières années. Le graphe distribué résultant des entités interconnectées sur le web est communément appelé le nuage LOD (voir figure 1.8) et couvre de nombreuses sources de données fournissant plus de 30 milliards de triplets RDF (Alhajj et Rokne, 2018). Ainsi, LOD couvre divers domaines allant des contenus liés aux médias, aux réseaux sociaux, aux contenus produits par les utilisateurs, aux données bibliographiques, à des données en sciences de la vie, en médecine, en biologie, à des données géographiques et des données gouvernementales. De plus, certaines sources de données telles que DBPedia [16] et YAGO [17] fournissent des informations générales interdomaines et jouent ainsi un rôle central dans la connexion de données provenant de domaines très différents.

1.5 Outils du web sémantique pour l'organisation de la connaissance en e-santé

Les organismes de santé du monde entier sont confrontés à des pressions pour réduire les coûts, améliorer la coordination et les résultats, et fournir plus avec moins. Cela nécessite des approches de planification efficace et une pratique fondée sur des preuves en générant des informations à partir des données disponibles. Ces dernières années, l'intégration des techniques du web sémantique a permis de nombreuses avancées, no-

14. Resoure Desription Framework attributes
15. Semantic Web Education and Outreach
16. `https://wiki.dbpedia.org/`
17. `https://github.com/yago-naga/yago3`

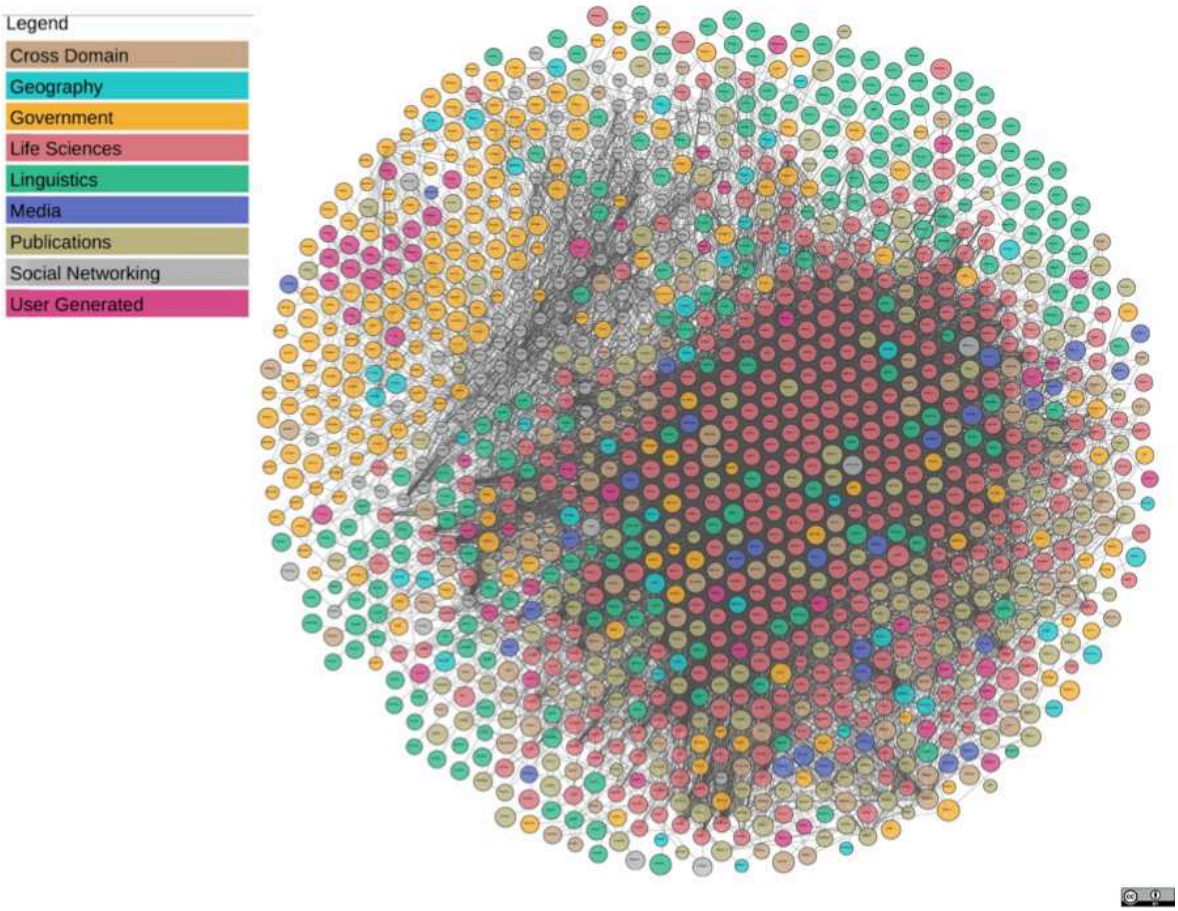


FIGURE 1.8 – Nuage du web de données (source : `https://lod-cloud.net/`)

tamment pour l'encodage, le stockage et la réutilisation des connaissances médicales. Le but de cette section est d'introduire un bref panorama de ces avancées, en présentant en particulier les thésaurus et les ontologies dans le domaine de la santé (section 1.5.1) et quelques référentiels de données sémantiques (section 1.5.2) en rapport avec notre domaine d'application.

1.5.1 Thésaurus et ontologies

Les thésaurus et les ressources termino-ontologiques sont devenus des outils importants dans le domaine de la santé, utilisés aussi bien dans le cadre des soins au patient que dans celui de la recherche.

Au cours des dernières années, plusieurs thésaurus et ontologies du domaine de la santé ont été créés. La plupart de ces ressources ont été créées pour définir un vocabulaire associé à une discipline spécifique médicale, tels que les termes décrivant les parties anatomiques et leurs relations, ou les termes utilisés en médecine clinique, tels que les dossiers de santé électroniques, etc. La plupart d'entre elles sont disponibles *via* BioPortal (BIOPORTAL, 2018), AgroPortal (JONQUET *et al.*, 2018), Linked Open Vocabularies (LOV, 2018).

1.5.1.1 Thésaurus

Un thésaurus est un ensemble structuré de termes normalisés organisés au sein d'une hiérarchie de concepts liés par des relations sémantiques. Les termes sélectionnés

sont nommés concepts ou descripteurs parce qu'ils sont destinés à l'analyse de contenu et au classement des documents. La plupart de ces thésaurus ont connu une série d'améliorations pour répondre à des besoins spécifiques. Ils sont d'abord été déployé au moyen de bases de données relationnelle mais la communauté de gestion a vite compris les limites de cette approche. La limite été associée à la nature distribuée des contributeurs et la disponibilité des données à des tiers uniquement au moyen de bases de données ou de services web. Un tel processus de partage des informations imposait un lourde contribution aux développeurs et aux responsables de la maintenance des applications. Afin de remédier à cette lacune,ils se sont tournés vers les données liées qui offraient un modèle standard lisible par l'homme et la machine.

Les thésaurus sont souvent représentés en SKOS (*Simple Knowledge Organization System*). SKOS est une recommandation du W3C pour représenter des vocabulaires contrôlés (Miles *et al.*, 2005). Ce standard, exprimé en OWL a pour principal objectif de faciliter la publication de vocabulaires structurés pour leur utilisation dans le cadre du web sémantique (Vandenbussche, 2011). Il définit une classe (*skos :Concept*) pour représenter des concepts, auxquels peuvent être associés un ou plusieurs termes (*skos :label, skos :preferedLabel, skos :hiddenLabel*). Le langage SKOS pemet également d'autres propriétés (*skos :definition, skos :broader, skos :narrower, skos :related*) permettant d'étendre son usage à d'autres types de modèles (glossaires, taxonomies, thésaurus, etc.).

AGROVOC : est un vocabulaire contrôlé (Caracciolo *et al.*, 2013) couvrant tous les domaines d'intérêt de l'Organisation des Nations Unies pour l'alimentation et l'agriculture (FAO), y compris l'alimentation, la nutrition, l'agriculture, la pêche, la foresterie, l'environnement, etc. Il est publié par la FAO et édité par une communauté d'experts. Il est disponible sous forme de schéma conceptuel SKOS-XL, également publié sous la forme d'un ensemble de données liées comprenant plus de 35 000 concepts disponibles dans 29 langues. Les termes de vocabulaire structurés et contrôlés d'AGROVOC permettent une identification non ambiguë des ressources, une normalisation des processus d'indexation et une recherche efficace.

LanguaL : est un thésaurus multilingue des aliments (Møller et Ireland, 2018) utilisant une classification à facettes. Chaque aliment est décrit par un ensemble de termes standards contrôlés choisis parmi les facettes caractéristiques de la qualité nutritionnelle ou hygiénique d'un aliment, telles que l'origine biologique, les méthodes de cuisson et de conservation, et les traitements technologiques. Le travail sur LanguaL a été lancé à la fin des années 1970 par le Centre pour la sécurité alimentaire et la nutrition appliquée (CFSAN) de la Food and Drug Administration (FDA) américaine, dans le cadre d'un effort de coopération permanent de spécialistes des technologies de l'alimentation, des sciences de l'information et de la nutrition. Depuis lors, LanguaL a été développé en collaboration avec le National Cancer Institute (NCI) des États-Unis et, plus récemment, ses partenaires européens, notamment en France, au Danemark, en Suisse et en Hongrie. Depuis 1996, le comité technique européen LanguaL administre le thésaurus. Un problème concernant les thésaurus multilingues est la multiplicité des langues : les termes correspondants de différentes langues ne sont pas toujours sémantiquement équivalents. Chaque descripteur est identifié par un code unique pointant vers des termes équivalents dans différentes langues (par exemple tchèque, danois, anglais, français, allemand, italien, portugais, espagnol et hongrois). Il facilite ainsi les liens vers de nombreuses banques de données sur les aliments et contribue à un échange de données cohérent.

MeSH [18] **:** est un thésaurus conçu pour l'indexation de publications médicales et la recherche d'information pour les sciences de la vie (NCBI, 2018). Les concepts y sont classés de manière hiérarchique selon 16 axes. La hiérarchie de concepts est organisée de manière à ce que tous les documents indexés par un concept soient également pertinents pour tous ses concepts parents. Les concepts sont décrits par des définitions textuelles, des attributs (exprimant par exemple des relations de synonymie) et des « qualifiers » permettant de préciser le contexte (par exemple « épidémiologie ») (Meilender, 2013).

1.5.1.2 Ontologies

Les ontologies sont des ressources importantes dans le domaine de la santé qui servent de composant central dans l'infrastructure des connaissances du web sémantique. La définition la plus consensuelle d'une ontologie est celle proposée par Studer [19] (Studer *et al.*, 1998) : une ontologie est une spécification formelle explicite d'une conceptualisation partagée d'un domaine donné. *formelle* exprime que la spécification doit être lisible par une machine, *explicite* signifie que les types des concepts et les contraintes sur leur utilisation sont explicitement définis, *conceptualisation* fait référence à un modèle abstrait d'un phénomène du monde reposant sur l'identification des concepts pertinents le caractérisant et *partagée* se rapporte à une construction consensuelle de la connaissance, qui n'est pas propre à un individu, mais validée par un groupe.

Les ontologies fournissent des vocabulaires contrôlés de la terminologie scientifique utilisés pour aider à l'annotation des données produites, tels que les termes et relations de base dans un domaine d'intérêt, ainsi que des règles sur la manière de combiner ces termes et relations. Elles sont utilisées pour fournir une terminologie commune à un domaine et la base de l'interopérabilité des systèmes.

PIPS [20] **:** est un projet dans le domaine de la santé (Cantais *et al.*, 2005) en ligne financé par la Commission européenne dans le cadre de l'appel 6. Ce projet visant à créer de nouveaux modèles de prestation de soins de santé en créant un environnement favorable à la santé et au soutien des services de la connaissance. Cet environnement intègre différentes technologies afin de permettre : (i) aux professionnels de la santé d'avoir accès à des connaissances médicales pertinentes et actualisées ; (ii) aux citoyens européens de choisir un mode de vie plus sain. Le projet vise à réunir les fournisseurs de soins de santé, les citoyens, les organisations publiques, les industries et services alimentaires ou pharmaceutiques, les chercheurs et les décideurs du secteur de la santé afin de créer un environnement de connaissances dynamique. Cet environnement dynamique s'appuie sur des approches traditionnelles et nouvelles de gestion des connaissances issues de la pratique médicale actuelle, de la médecine factuelle et de sources de connaissances disparates dans les domaines de la santé ou nutrition.

FOODS : est un système (Snae et Bruckner, 2008) de conseil pour la planification des aliments ou des menus dans un restaurant, une clinique, un hôpital ou à domicile. Le système est piloté par une ontologie axée sur l'alimentation portant le même nom que le système (FOODS). FOODS comprend : (a) une ontologie alimentaire ; (b) un système expert utilisant cette ontologie et des connaissances sur les méthodes de cuisson et les prix ; (c) une interface utilisateur adaptée aux novices en

18. Medial Subjet Headings
19. « *An ontology is a formal, explicit specification of a shared conceptualization* »
20. Plateforme d'informations personnalisée pour la santé et les services de la vie

informatique et en régimes ainsi qu'aux experts. L'ontologie contient les spécifications des ingrédients, des substances, des informations nutritionnelles, des doses journalières recommandées pour différentes régions, plats et menus.

Food Product Ontology : est une ontologie (Kolchin et Zamula, 2013) décrivant une langue commune relative à la nutrition des produits alimentaires en utilisant un vocabulaire exprimé dans le domaine des produits alimentaires. L'objectif est d'aider les fabricants, les détaillants, les gouvernements et les institutions à publier leurs données relatives à ce domaine de manière à maximiser la réutilisation des données. L'ontologie des produits alimentaires permet une meilleure intégration, le partage et le traitement collaboratif des informations sur les aliments entre plusieurs parties prenantes. Il étend une ontologie normalisée largement utilisée pour les données sur les produits, les prix, les magasins et les sociétés appelée GoodRelations (Hepp, 2008), qui est une ontologie décrivant des biens matériels et des services de marchandises utilisant des données structurées en format RDF et microformats.

FoodWiki : est une ontologie (Çelik, 2015) intégrée dans un système de santé mobile en ligne fondé sur une consommation alimentaire adaptée. Il permet aux utilisateurs ou patients de groupes à risque de surveiller et de contrôler leur consommation alimentaire. Plus précisément, les patients seront en mesure d'éviter les ingrédients malsains susceptibles d'aggraver leur état de santé. Le système de santé en ligne mobile est doté d'une base de connaissances et d'un mécanisme de recherche qui prend en compte les requêtes des utilisateurs et suggère une consommation alimentaire appropriée et adaptée à chaque individu. FoodWiki considère uniquement les produits alimentaires emballés dans les rayons du marché.

FoodOn : est un projet mené par un consortium visant à créer une ontologie mondiale complète et facilement accessible de la ferme à l'assiette, qui décrit de manière précise et cohérente les aliments connus dans les cultures du monde entier (FoodOn, 2018). La première version de FoodOn est basée en grande partie sur LanguaL. Les aliments peuvent également être liés à leurs origines culturelles ou géographiques et à leur taxonomie des matières premières animales, végétales ou fongiques. La définition logique de produits alimentaires est introduite *via* la base de données LanguaL-indexed SIREN [21]. Actuellement, ces produits sont organisés manuellement dans la hiérarchie de la facette « foodon, food, product », mais celle-ci sera transformée en une structure déduite à mesure que les définitions de produits sont mises à jour. L'objectif est d'exploiter les relations de l'ontologie pour permettre d'améliorer la façon dont les composants et les processus sont représentés afin de tirer parti des avantages du raisonnement automatisé sur les schémas de produits alimentaires.

SNOMED CT : a été développée en collaboration entre pays pour s'assurer qu'elle répond aux divers besoins et attentes des cliniciens du monde entier. Elle est désormais acceptée comme un langage mondial commun pour les termes de santé et considérée comme la principale ontologie de représentation des concepts cliniques, des termes et des relations dans le domaine des soins de santé (SNOMED-CT, 2018). Le travail de codage et de classification a été guidé par trois critères de conception fondamentaux qui sont la compréhension, la reproductibilité et l'utilité. L'ontologie est structurée selon une hiérarchie avec un ensemble de concepts généraux de haut niveau. Tous les autres concepts sont des sous-types de ces concepts principaux, et plus nous évoluons dans la hiérarchie, plus les concepts deviennent spécialisés. SNOMED CT couvre la plupart des domaines utilisés dans la pratique médicale, y compris les signes cliniques, les symptômes, les diagnostics, les produits pharmaceutiques, les structures corporelles,

21. `http://www.langual.org/langual_indexed_datasets.asp`

les dispositifs médicaux, les contextes sociaux, etc. Chaque concept se voit attribuer un identifiant unique et un nom entièrement spécifié, ce qui peut être interprété comme une description lisible pour l'humain.

D'autres ontologies spécifiques concernant certaines maladies ont également été développées et indexées dans la catégorie BioPortal sous la rubrique santé. L'ontologie développée par (Min *et al.*, 2009) pour aider à la gestion et l'intégration des données cliniques sur le cancer de la prostate. L'ontologie CVDO sur les maladies cardiovasculaires qui réorganise et complète les maladies cardio-vasculaires autour du modèle de maladie tripartite des OGMS, et construit sa hiérarchie des maladies en grande partie à l'aide d'un raisonnement automatique (Barton *et al.*, 2014). Des ontologies relatives aux maladies rares telles que l'ontologie de la maladie d'Alzheimer (Malhotra *et al.*, 2014) sont également présentes.

La question de l'utilisation des ontologies médicales est traitée dans la littérature, (Fung et Bodenreider, 2012) ont identifié deux principales difficultés. La première est le nombre important d'ontologies médiales. Plusieurs centaines sont disponibles, sans qu'il existe de métrique communément admise pour leur validation et leur certification (Bodenreider, 2006). La seconde difficulté est le manque d'interopérabilité entre ces ontologies. Ainsi, (Bodenreider, 2006) montrent que l'alignement du concept « blood » entre différentes ontologies de l'UMLS mène à des incohérences.

1.5.1.3 Bases de données, terminologies et classifications

Open Food Facts : est une base de données (Open-Food-Facts, 2018) sur les produits alimentaires construite de manière collaborative et accessible à tous publics dans le monde. Elle est disponible sous différents format (CSV, RDF, etc.). En octobre 2018, la base de données Open Food Facts couvrait 659752 produits alimentaires provenant du monde entier et 416761 produits de France. Les utilisateurs ajoutent des produits à la base de données et ont le privilège de modifier, d'améliorer et de proposer des idées d'applications *via* le forum. L'application Open Food Facts permet aux utilisateurs de se renseigner sur les informations nutritionnelles des aliments et de comparer des produits du monde entier. Cela se fait simplement en scannant le code à barres du produit alimentaire ou en effectuant une recherche dans la base de données. L'objectif est de connecter les utilisateurs avec des outils intelligents pour les aider à faire de meilleurs choix en matière de consommation alimentaire. Les produits alimentaires sont étiquetés selon le Nutri-Score.

MedDRA [22] **:** une terminologie médicale standardisée riche et hautement spécifique facilitant le partage d'informations réglementaires internationales sur les produits médicaux utilisés par l'homme (MedDRA, 2018). Les produits couverts par MedDRA incluent les produits pharmaceutiques, les produits biologiques, les vaccins et les produits combinés. MedDRA est à la disposition de tous pour l'enregistrement, la documentation et la surveillance de la sécurité des produits médicaux avant et après la mise en vente d'un produit.

CIM-10 [23] **:** est publiée par l'OMS en tant que système de classification des maladies et autres aspects connexes de la pratique médicale : symptômes, anomalies, causes des maladies, certificats de décès et dossiers médicaux. La CIM-10 est utilisée par les États membres de l'OMS pour la compilation de statistiques nationales sur la mortalité et la morbidité, pour la recherche épidémiologique et pour l'évaluation des tendances

22. Dictionnaire médical pour les activités de réglementation
23. Classification internationale des maladies, dixième révision

en matière de santé publique et de maladies. La CIM-10 n'est pas une ontologie au sens strict, mais elle est généralement exploitée comme thésaurus et classification. Ces dernières années, une version OWL de la CIM-10 a été développée pour fournir une manière plus formelle et plus efficace de décrire les termes et leurs relations.

1.5.1.4 Synthèse et discussion

La place des ontologies et des thésaurus dans les systèmes de santé est très importante pour pallier l'ambiguïté d'interprétation, favoriser l'échange d'une information univoque et ainsi contribuer de manière plus générale à l'interopérabilité sémantique. Comme nous l'avons introduit précédemment, il existe des terminologies plus ou moins formalisées, des thésaurus avec une structure plus ou moins complexe, des ontologies avec ou sans contraintes sur leurs relations. La plupart des modèles de connaissances traités *supra* proposent des relations permettant d'organiser les concepts dans une hiérarchie en ajoutant des liens sémantiques entre eux. La profondeur de ces hiérarchies varie et dépend principalement du niveau de granularité souhaité. Un caractère commun aux quelques exemples présentés *supra* est leur capacité à gérer le langage naturel pour les humains leur facilitant ainsi leurs maintenances et leurs utilisations.

Ces ressources sont généralement construites selon un seul point de vue ce qui limite parfois leur réutilisation. Dans le cas des aliments par exemple, il existe de nombreuses classifications construites selon des critères qui ne sont pas forcément pertinents pour notre travail. Il s'agit par exemple de la classification botanique, des différents lexiques associés au site de cuisine, des dictionnaires et plus particulièrement celui de la cuisine, des thésaurus existants dans le domaine de la nutrition. Or, les aliments peuvent être classés selon plusieurs points de vue en fonction des usages relevant des différents domaines (domaine sensoriel, domaine de la cuisine, domaine de la nutrition, etc.). Par exemple, si nous cherchons à faire de la qualification nutritionnelle de recettes de cuisine faites maison, la ressource Open Food Facts ne sera pas totalement exploitée. En effet, elle décrit des produits manufacturés or dans la préparation de recettes il faut considérer des aliments qui vont subir différents traitements pour aboutir au plat final. Dans ce cas de figure, l'appel à d'autres ressources est nécessaire avec la mise en place d'approches permettant l'intégration de ces différentes ressources.

1.5.2 Référentiels de données sémantiques

Ces dernières années, il y a eu une explosion des données sur la santé en termes d'hétérogénéité et de volume. La plupart de ces données sont publiées dans des formats propriétaires, et certains de ces formats sont conformes au bonnes pratiques recommandées par le W3C, permettant ainsi un accès *via* HTTP et en mettant l'accent sur les ensembles de données pour établir des relations et des interconnexions entre différentes sources de données.

(Darmoni *et al.*, 2000) proposent CISMeF un service du même type à destination des professionnels de santé. Il indexe les ressources médicales francophones disponibles en ligne. Chaque site web est indexé selon une notice descriptive reposant principalement sur MeSH. Cette indexation tend à devenir multi-terminologique. CISMeF dispose de son propre moteur de recherche proposant des fonctionnalités analogues à celles du moteur de PubMed.

CardioSHARE (Vandervalk *et al.*, 2008) représente une infrastructure de services web décentralisée qui fournit un point de terminaison SPARQL permettant d'interroger de manière transparente des sources distribuées et indépendantes. Bien que l'objectif

principal du projet soit l'analyse de données cliniques sur les maladies cardiaques, le cadre est générique et peut être appliqué à tous types de données. Le système est évolutif et, plus important encore, il fournit un accès uniforme aux données distribuées sous format RDF à partir des formats et des propriétaires différents.

(Belleau *et al.*, 2008) proposent Bio2RDF, un projet open source qui utilise les technologies du web sémantique pour fournir l'un des plus grands réseaux de données liées appartenant à un ensemble diversifié de sources au format hétérogène et obtenues auprès de plusieurs fournisseurs de données. Le système fournit un point de terminaison SQPRQL qui peut être utilisé pour interroger environ 11 milliards de triplets provenant de 5 ensembles de données différents issus d'essais cliniques, de PubMed et d'autres grands fournisseurs de données du domaine biomédical.

OntoWiki (Zaveri *et al.*, 2013) est une plateforme qui fournit un point de terminaison SPARQL pour interroger des données de l'OMS. L'ensemble de données contient des millions de triplets. Il est publié sous forme de données liées. OntoWiki permet également de parcourir les données au format HTML.

HCLS (Web in Health Care and Life Sciences) (Marshall et Prud'hommeaux, 2008) est une base de connaissances biomédicales. Elle est produite par le groupe qui fait partie de l'activité web sémantique du W3C et intègre 15 sources de données distinctes utilisant les technologies web (RDF, OWL, etc.). Bien que l'utilité de la base de connaissances soit illustrée par l'identification d'un ensemble de gènes impliqués dans la maladie d'Alzheimer, l'approche décrite peut être généralisée à tous les cas d'utilisation intégrant des données provenant de plusieurs domaines.

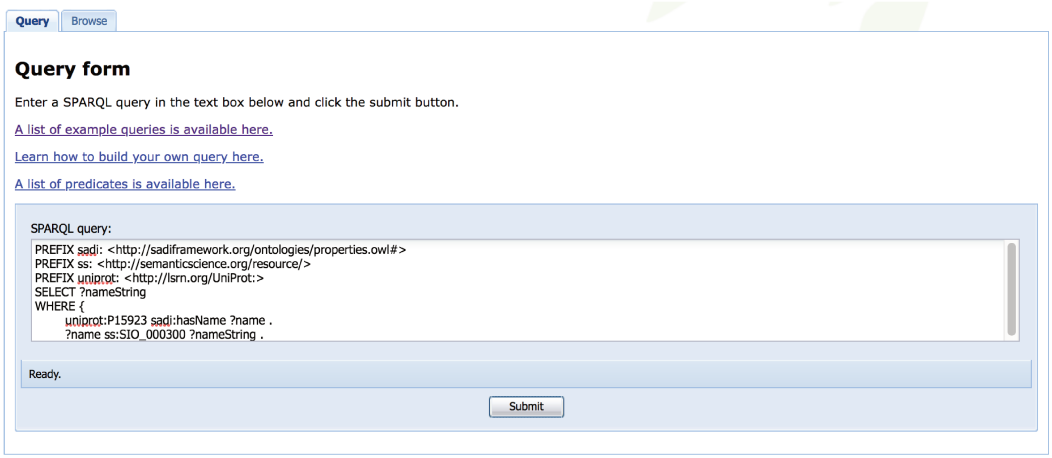


Figure 1.9 – Interface de CardioShare

Les projets utilisés pour produire des données sémantiques dans le domaine de la santé, tel que CardioShare, LinkedCT, Bio2RDF et autres, offrent déjà des fonctionnalités de navigation et de recherche dans les données sémantiques sous-jacentes. Cependant, leurs interfaces actuelles sont limitées et non intuitives. Les utilisateurs qui ne maîtrisent pas les technologies du web sémantique peuvent avoir des difficultés à rechercher, naviguer, afficher et visualiser les triplets RDF. Par exemple, l'interface CardiShare offre des fonctionnalités de navigation, un point de terminaison SPARQL et un formulaire de requête. Cependant, la recherche de données nécessite l'écriture de requêtes au format SPARQL, comme illustré à la figure 1.9, qui nécessite des connaissances spécifiques sur les données sous-jacentes et les spécificités du langage de requête SPARQL. Dans le cas des maladies cardiovasculaires, les ontologies décrivant les maladies cardiovasculaires sont incomplètes et ne décrivent pas l'ensemble des termes pour

pouvoir faire de la prédiction. Un effort de modélisation des connaissances du domaine est par conséquent nécessaire.

1.6 Conclusion partielle

Le projet proposé dans cette thèse s'inscrit dans le cadre de l'amélioration de la prévention des maladies cardiovasculaires. Plus précisément, son but est de fournir des outils et des méthodes pour l'évaluation du risque cardiovasculaire, fournir des recommandations adaptées au profil de l'individu en exploitant les connaissances du domaine. Afin d'y parvenir, de récents travaux ont mis en avant le besoin d'une évaluation du risque cardiovasculaire en prenant en compte les interactions entre les facteurs de risque cardiovasculaire. Le cadre proposé dans ce travail permet d'évaluer le risque en s'appuyant sur les recommandations de la HAS et des connaissances expertes en mettant en place des techniques et des outils fondés sur le web sémantique, un ensemble de standards permettant la formalisation et l'exploitation automatique des connaissances. Dans le chapitre suivant, la solution proposée pour modéliser les connaissances sur les facteurs de risque cardiovasculaire est présentée. Cette modélisation permet une meilleure compréhension des interactions existant entre les facteurs de risque et par conséquent une meilleure évaluation du risque. Dans la suite de ce travail, nous montrons de quelle manière ces facteurs peuvent être utilisés dans le cadre de l'évaluation et de la prévention des maladies cardiovasculaires.

Chapitre 2

Exploitation des présentations de résultats statistiques : application à la modélisation des facteurs de risque cardiovasculaire

Sommaire

2.1 Définition et terminologie **40**
2.1.1 Statistiques 40
2.1.2 Démarche statistique 40
2.1.3 Présentation des résultats statistiques 41
2.1.4 Discussion 42
2.2 Approches d'extraction de connaissances à partir de textes **43**
2.2.1 Approche sur textes non structurés 43
2.2.2 Approche sur textes structurés 45
2.2.3 Discussion et positionnement 50
2.3 Approche proposée **50**
2.3.1 Conversion des documents de présentation de résultats statistique 50
2.3.2 Localisation et extraction de l'information 52
2.3.3 Construction des triplets 54
2.4 Expérimentation **56**
2.4.1 Domaine des maladies cardiovasculaires 56
2.4.2 Domaine financier 59
2.5 Evaluation de l'approche **60**
2.6 Conclusion **64**

Ce chapitre vise à fournir une approche d'extraction automatique des connaissances à partir de textes. Dans notre étude, nous nous intéressons plus précisément à l'extraction et l'interprétation de connaissances à partir de tableaux statistiques publiés au format PDF. Nous commençons par décrire la démarche statistique d'interprétation des résultats et les méthodes visuelles les plus utilisées pour les présenter. Ensuite, nous présentons et comparons les approches d'extraction des connaissances à partir de ces présentations avant d'exposer notre approche. Nous expérimentons ensuite l'approche d'extraction proposée sur deux jeux de données relatives aux domaines des maladies cardiovasculaires et de la finance. Enfin, l'approche est évaluée.

Texte de référence :
Rabia Azzi, Sylvie Despres, Jérôme Nobécourt. D'un modèle statistique à un modèle de connaissance : retour d'expérience. Sylvie Ranwez. 29es Journées Francophones d'Ingénierie des Connaissances, IC 2018, pp.105-119, 2018 Nancy, France.

2.1 Définition et terminologie

2.1.1 Statistiques

L'histoire des statistiques en tant que discipline reconnue intervient en 1900, l'année du khi-carré [1] de Karl Pearson. Avant cela, les statistiques étaient proches du monde de Quetelet, où d'énormes ensembles de données de recensement étaient utilisés pour répondre à des questions simples, mais importantes (par exemple, y a-t-il plus de naissances masculines ou féminines ? est-ce que le taux de meurtre augmente ? etc.) (Lovric, 2011). Cette discipline s'est largement développée et joue un rôle de premier plan dans la plupart des branches de la recherche scientifique, qu'il s'agisse d'expériences de laboratoire bien contrôlées ou de grandes enquêtes (Maxwell, 1978).

Selon (Iversen et Gergen, 1997), le terme statistique possède de nombreuses significations, certaines d'entre elles étant mieux définies que d'autres. Le mot statistique lui-même semble avoir été inventé par un allemand nommé Hermann Conring quand il introduit le terme Statistik en 1660. La première partie du mot est une adaptation du mot état (Staat en allemand), qui correspond au terme utilisé pour nommer la pratique des États recueillant des informations sur les naissances et les décès il y a plus de trois cents ans.

La statistique est un ensemble de concepts, de règles et de méthodes pour : (i) collecter des données (ii) analyser des données et (iii) tirer des conclusions à partir de données (Wolfgang *et al.*, 2015). 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.).

Dans la suite de ce document, nous adopterons la définition suivante : « La statistique est le domaine d'étude dont l'objectif est l'exploitation des données (généralement des ensembles de chiffres et des caractéristiques d'identification) conduisant à des interprétations sur ces données (habituellement présentés sous forme de tableaux, de graphiques et de résumés écrits et verbaux) ».

2.1.2 Démarche statistique

La démarche statistique comporte usuellement les étapes suivantes (Saporta, 2011) :

1. est un test statistique où la statistique de test suit une loi du χ^2 sous l'hypothèse nulle.

— 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 et structurer l'information contenue dans les données ;

— l'inférence, qui consiste à étendre les propriétés constatées sur l'échantillon de la population tout 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.

2.1.3 Présentation des résultats statistiques

Les méthodes visuelles sont largement utilisées en statistique pour présenter les résultats de manière claire et concise. Selon (Heiberger et Holland, 2015 ; Junyong et Sangseok, 2017), il existe plusieurs formes possibles de présentations visuelles des résultats : tableaux, graphiques, histogrammes, diagrammes, etc. La présentation textuelle est la principale méthode pour expliquer les résultats, décrire les tendances et fournir des informations contextuelles. Un tableau est le mieux adapté pour présenter des informations individuelles, quantitatives et qualitatives. Un graphique est un outil visuel très efficace, car il permet d'appréhender les données en un coup d'œil, de faciliter la comparaison et de révéler des relations au sein des données.

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 2.1) est d'adopter une démarche allant du général au plus détaillé. Cette démarche comporte les étapes suivantes :

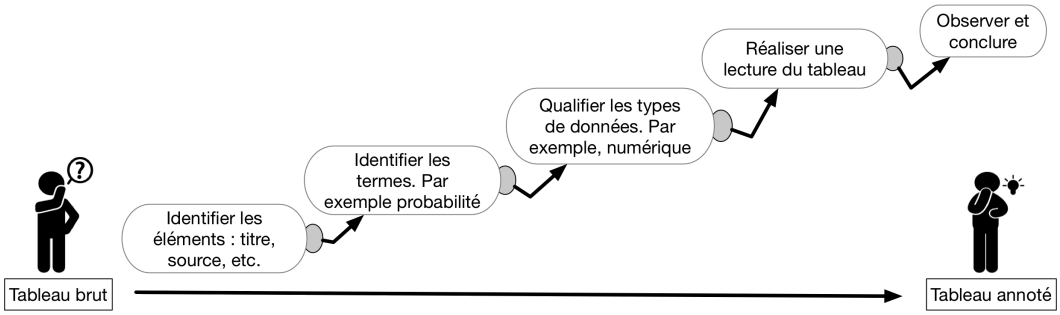


Figure 2.1 – Démarche d'analyse d'un tableau statistique

— identifier les éléments : identification du titre, de la source de l'étude, de 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 : identification des termes figurant dans le titre, les colonnes et les lignes du tableau, etc. ;

— qualifier les types de données : qualification du type de données contenu dans le tableau en prenant en compte les unités (par exemple, les pourcentages, les probabilités, etc.) ;

— réaliser une lecture du tableau : application de 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 : conclusion tirée à partir du tableau. Par exemple, identifier des relations entre variables (cause/effet), des valeurs extrêmes, des tendances, etc.

Un tableau statistique 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. Dans ce cas, il existe trois types de tableaux : (i) les tableaux de données (les premiers construits) 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.

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 et Kyong-Ho, 2008), il existe deux types de tableaux (voir figure 2.2) : (1) les tableaux simples (voir figure 2.2-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 2.2-B) sont constitués d'en-têtes composés 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.

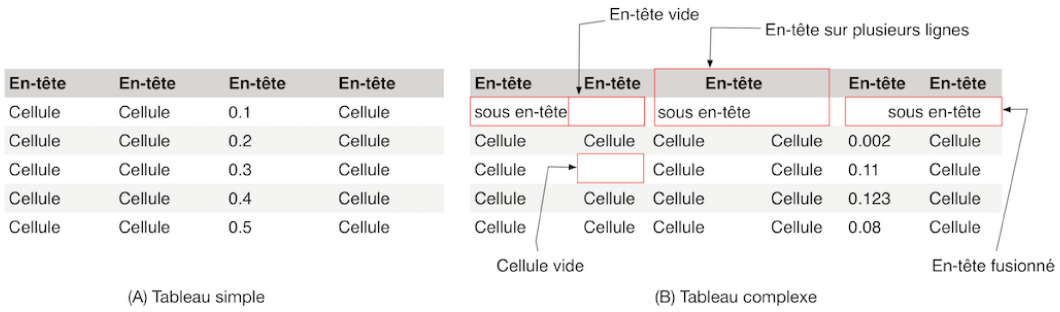


Figure 2.2 – Exemple de tableau simple et complexe

2.1.4 Discussion

La section précédente a mis en avant la démarche statistique et l'approche d'analyse des tableaux qui est à la base de l'interprétation des informations relationnelles dans les tableaux. L'objectif de ce travail est d'automatiser cette démarche analytique pour extraire les informations et les relations contenues dans ces tableaux. Les informations représentées dans les tableaux sont fortement connectées. Les liens entre les éléments d'un tableau sont implicites et nécessitent une explicitation. Cependant, comme toutes les données sont d'égale importance, il n'est pas facile d'identifier et de sélectionner les informations essentielles.

Si une analyse visuelle permet à un être humain de reconnaître et comprendre facilement les tableaux, la situation est différente quand il s'agit d'un ordinateur. Considérons à titre d'exemple, le tableau (B) de la figure 2.2, 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 de type différents (texte, chiffres, caractères spéciaux, etc.).

En règle générale, 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.

Pour réduire ce biais d'interprétation lors de l'automatisation du processus, nous proposons d'extraire et de structurer les connaissances implicites contenues dans ces tableaux sous forme d'un modèle conceptuel. L'avantage d'une telle représentation est : (i) d'obtenir une description structurée explicitant les connaissances du domaine ; (ii) d'être souple en permettant d'étendre le modèle à d'autres concepts du domaine.

2.2 Approches d'extraction de connaissances à partir de textes

Au cours des 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). Celles-ci partagent un même objectif final, celui du traitement automatique des textes, mais se différencient sur plusieurs points.

Un *texte* peut être vu comme une suite de mots (ou *token*) séparés par des espaces et un ensemble de caractères de ponctuation. Les mots sont groupés dans des phrases, elles-mêmes groupées en paragraphes. Une séquence de paragraphes constitue une section, une suite de sections constitue un texte. L'extraction des connaissances à partir du texte peut être réalisée selon trois approches (voir figure 2.3) :

1. les approches symboliques sont fondées sur l'hypothèse qu'il est possible d'exploiter la grammaire de structuration pour la mise en forme ou la mise en page des textes ;
2. les approches statistiques sont fondées sur l'hypothèse qu'il est possible d'apprendre à partir de régularités et de les généraliser à de nouvelles données. ;
3. les approches hybrides sont fondées sur l'utilisation des approches symboliques associées à des approches statistiques.

2.2.1 Approche sur textes non structurés

L'extraction de connaissances à partir de textes non structurés consiste à mettre en place des approches où les éléments de mise en forme ne sont pas pris en compte.

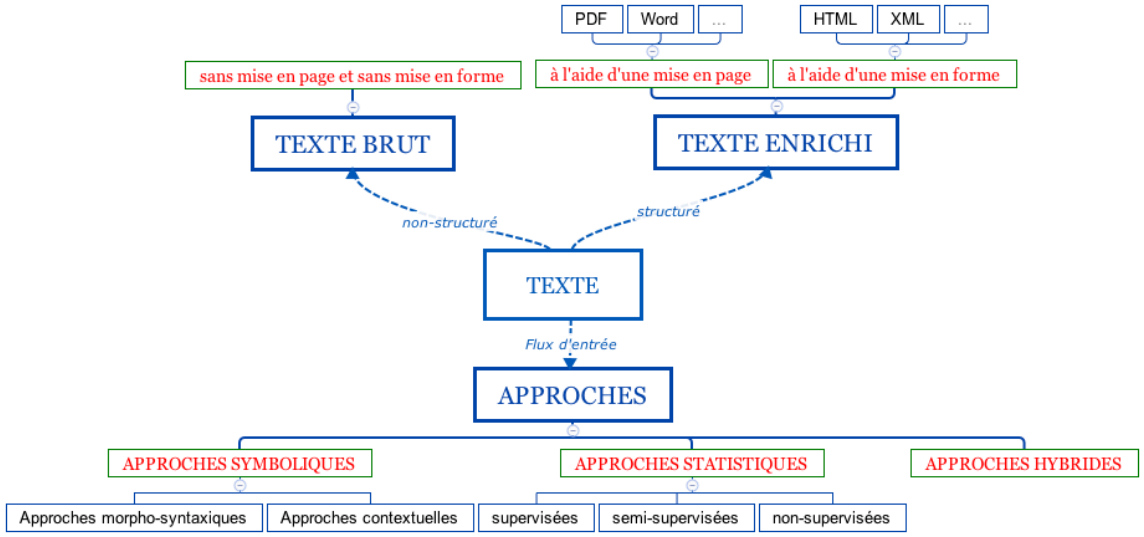


FIGURE 2.3 – Différentes approches d'extraction des connaissances à partir de texte

Les travaux de (GRABAR et ZWEIGENBAUM, 1999) dans le domaine médical s'intéressent aux connaissances que nous pouvons acquérir à partir d'un texte en considérant la structure morphologique du texte pour l'extraction de relations sémantiques. Dans leur approche, ils exploitent une liste de termes synonymes disponibles afin d'amorcer le processus d'acquisition. Il considère comme *morphologiquement* reliés une paire de mots qui sont dérivés de la même racine commune et, donc, partagent plus ou moins un sens commun. Dans le même contexte, (ZWEIGENBAUM *et al.*, 2003) proposent une approche fondée sur la cooccurrence en corpus de mots formellement proche et un filtrage complémentaire sur la forme des mots dérivés.

Toujours dans ce domaine de l'extraction de relations sémantiques, d'autres approches contextuelles ont été proposées. (M. HEARST, 1992) propose de s'appuyer sur des patrons lexico-syntaxiques pour identifier des relations lexicales d'hyponymie à partir de textes. L'approche consiste à identifier des schémas lexico-syntaxiques facilement reconnaissables qui se produisent de manière régulière et qui indiquent indiscutablement la relation lexicale d'intérêt. De façon analogue, (AUSSENAC-GILLES et SÉGUÉLA, 2000) proposent l'extraction et la modélisation de relations sémantiques à partir de corpus textuels. (ARNOLD et RAHM, 2014) présentent une approche fondée sur des patrons sémantiques pour identifier les relations sémantiques entre les concepts. L'approche est appliquée sur des articles Wikipedia pour extraire plusieurs relations à partir d'un seul article. Un autre exemple d'approche proposé par (RAMADIER et LAFOURCADE, 2016) consiste à extraire des relations sémantiques dans les textes médicaux et plus particulièrement dans les comptes rendus radiologiques. Cette approche est fondée sur l'utilisation de patrons sémantiques vérifiant des contraintes dans une base de connaissances.

Dans le domaine de la construction de base de connaissances, (TOUNSI *et al.*, 2017) proposent une approche semi-supervisée pour extraire des relations à partir de textes dans le domaine du cosmétique. Dans le même ordre d'idées, (JILANI et AMARDEILH, 2009) proposent de construire une ontologie et de la peupler automatiquement grâce à une méthode de traitement automatique des langues fondée sur les patrons lexico-syntaxiques. D'autres approches non-supervisées comme celle proposée par (RAJMAN et BESANÇON, 1998) consiste à extraire des connaissances à partir de textes non structuré

avec une approche probabiliste d'association entre les entités extraites à partir du texte.

(Nakamura-Delloye et Villemonte, 2010) proposent pour l'acquisition de patrons de relations entre entités nommées une approche semi-supervisée à partir de résultats d'analyse syntaxique. Sans aucun patron prédéfini, la méthode fournit des chemins syntaxiques susceptibles de représenter une relation donnée à partir de quelques exemples de couples d'entités nommées entretenant la relation en question.

(Nakamura-Delloye et Stern, 2011) proposent une méthode non-supervisée d'extraction des relations et des patrons de relations entre entités nommées, réalisées dans le cadre de la création et l'enrichissement d'une ontologie. La méthode proposée se caractérise par l'exploitation des résultats d'analyse syntaxique, notamment les chemins syntaxiques reliant deux entités nommées dans les arbres de dépendance. (Claveau et Ncibi, 2013) proposent une approche non-supervisée pour la découverte de connaissances. L'approche consiste à détourner les champs aléatoires conditionnels (CRF), qui ont montré leur intérêt pour des tâches d'étiquetages supervisées, pour calculer indirectement ces similarités sur des séquences de textes.

D'autres travaux comme (Malintha *et al.*, 2015) proposent un cadre qui sert de modèle d'apprentissage semi-supervisé pour obtenir des informations en établissant des relations entre différentes entités dans le contenu en ligne.

2.2.2 Approche sur textes structurés

L'extraction de connaissances à partir de textes structurés consiste à extraire des connaissances en utilisant les propriétés de formatage et de mise en page des documents. Deux catégories d'approches sont distinguées :

— les approches exploitant des formatages prédéfinis ;

— les approches exploitant des documents avec mise en page.

2.2.2.1 Approches sur des textes à formatages prédéfinis

Dans cette section, nous présentons les approches qui exploitent les **langages de balisage** (*markup languages*) pour l'extraction de connaissances. Nous regroupons ces approches selon le type de langage à balises sur lequel elles s'appuient : le format HTML, le format XML, etc.

Format HTML. Le format HTML (*HyperText Markup Language*) est un langage de balisage conçu pour la structuration des documents sur le web. Il se compose de deux parties essentielles : le contenu de l'information et un ensemble d'instructions (« balisage ») qui indique à l'ordinateur comment afficher ce contenu (Brooks, 2007). Syntaxiquement, le HTML hérite du format SGML (*Standard Generalized Markup Language*) auquel il ajoute néanmoins une sémantique prédéfinie sur un ensemble de balises (Berners-Lee et Connolly, 2003). HTML est devenu le langage du Web le plus utilisé. Voilà pourquoi diverses approches ont été proposées pour extraire les connaissances à partir de ces pages web en exploitant les balises HTML. Dans leurs travaux, (Shinzato et Torisawa, 2004a ; Shinzato et Torisawa, 2004b) proposent par exemple d'extraire des relations d'hyperonymie en utilisant les balises liste dans les pages HTML (par exemple). L'hypothèse est que les entités exprimées dans une même liste ont tendance à partager le même hyperonyme commun. D'autres approches non-supervisée proposent d'extraire des paires d'attributs-valeurs pour un objet donné à partir de documents HTML. (Yoshinaga et Torisawa, 2007) proposent d'extraire des

paires d'attributs-valeurs en deux phases. Premièrement, une acquisition des attributs pour une classe donnée (par exemple, « film »). Deuxièmement, en utilisant les attributs de la classe acquis, trouver des pages web susceptibles d'inclure les paires d'attributs-valeur de l'objet, puis de les extraires de ces pages en utilisant des expressions régulières induites automatiquement. (Ravi et Paşca, 2008) proposent une approche pour extraire des attributs de documents Web, en utilisant la structure existante (la hiérarchie de balises HTML) dans le texte. (Bounhas et Slimani, 2010) proposent d'exploiter la structure logique des documents HTML pour déduire les relations sémantiques entre les termes et construire ainsi une ressource termino-ontologique. L'approche correspond à l'indexation descendante qui accorde la plus grande importance aux termes apparaissant dans les nœuds principaux du document. Les termes sont pondérés en fonction de leur position dans la structure hiérarchique du document. Une fois les documents indexés, les relations logiques entre leurs fragments sont exploitées pour constituer un réseau contextuel de termes. Les liens de ce réseau permettent de déduire des relations sémantiques utiles pour l'organisation de la terminologie.

Actuellement, de grands volumes de données sont stockés sous forme de tableaux sur le web. Cette situation a conduit à de nombreux travaux centrés sur l'extraction de connaissances à partir de tableaux pour alimenter un grand nombre d'applications (Crestan et Pantel, 2010). Contrairement au texte non-structuré, les tableaux les lignes et les colonnes sont syntaxiquement et sémantiquement cohérentes. Cette structuration est bénéfique pour l'interprétation des données. La richesse et l'utilité des tableaux sur le web ont permis l'émergence de plusieurs approches d'extraction de connaissances. Par exemple, pour exploiter la sémantique des tables de Wikipédia et en extraire des faits, (Muñoz *et al.*, 2014) proposent une méthode d'extraction sous forme de triplets RDF à l'aide d'une base de connaissances « Linked Data » existante. L'approche consiste à trouver des relations préexistantes entre des entités dans les tables de Wikipédia, suggérant les mêmes relations que pour d'autres entités dans des colonnes analogues sur des lignes différentes. Dans le même ordre d'idées (Weiming *et al.*, 2015) proposent une approche pour l'extraction de triplets RDF à partir de tableaux contenus dans des encyclopédies chinoises. La méthode proposée consiste à : (i) construire une base de connaissances chinoise à travers deux processus, la construction d'une hiérarchie et l'extraction d'attributs de classe ; (ii) extraire à l'aide de la base de connaissances des triplets à partir des tables grâce à la notation par colonne, à la classification des tables et à l'extraction RDF.

D'autres approches se sont concentrées sur l'extraction des connaissances sous forme de paires attribut/valeur. Dans cette optique, (Mahesh *et al.*, 2012) proposent une approche fondée sur des règles pour traiter et extraire des paires d'attributs/valeurs et leur relation en se basant sur les cellules du tableau.

D'autres approches consistent à exploiter la structure logique des tableaux pour pouvoir interpréter leur sémantique. (Pivk *et al.*, 2007) proposent une méthodologie appelée TARTAR permettant de transformer et produire des trames sémantiques (F-logic) à partir de la structure et du type de tables. (Ding *et al.*, 2010) proposent quant à eux d'extraire des triplets RDF à partir de la structure des tables, en considérant les lignes des tables comme des sujets, les en-têtes des colonnes comme prédicats, et les valeurs des cellules comme des objets.

Format XML. Le format XML (*eXtensible Markup Language*) est le langage bien connu pour les documents et les données structurés. XML est une recommanda-

tion du W3C datant de 1998 inspirée de SGML (Fierz et Grütter, 2000). La première version de XML a introduit la distinction entre la structure physique et la structure logique à l'aide d'une DTD (*Document Type Definition* (Bray *et al.*, 1998)). XML est un langage de balises permettant essentiellement de décrire des arbres structurés dans un format textuel. Un document XML comporte des **éléments** [2] et des **attributs** [3] (Gandon *et al.*, 2012). La figure 2.4, présente un exemple de document XML constitué d'un élément RECETTE ayant trois éléments fils NOM, DUREE et REMARQUE ayant chacun un fils à valeur littérale. Un espace de noms est déclaré avec le préfixe « lr » associé à l'URI identifiant l'espace « http ://www-limics.smbh.univ-paris13.fr/recette# ».

```
<?xml version="1.0" encoding="UTF-8"?>

<lr:recette lr:id="12" xmlns:lr="http://www-limics.smbh.univ-paris13.fr/recette#">
    <lr:nom>Ratatouille</lr:nom>
    <lr:duree>1h20</lr:duree>
    <lr:remarque>Très bon chaud mais peut aussi se servir froid</lr:remarque>
</lr:recette>
```

Figure 2.4 – Exemple de document XML décrivant une recette de cuisine

L'extraction de connaissances exploitant la structure exprimée au format XML a donné lieu à de nombreux travaux ces dernières années. L'un des premiers workshops dédiés à cette tâche est le *Knowledge Discovery from XML Documents* (Nayak et J-Zaki, 2006). Dans ce workshop, (Brunzel et Spiliopoulou, 2006) ont présenté le système XTREEM (Xhtml TREE Mining) dédié à l'extraction de la sémantique des collections de documents Web, qui tire parti du contenu semi-structuré de XHTML [4]. Dans la perspective de créer des bases de connaissances, (Damiani *et al.*, 2004) ont proposé une approche fondée sur la structure des documents XML représentant des instances d'un modèle de données XML donné, appartenant à un domaine d'intérêt commun. Ils commencent par analyser le flux de données, pour obtenir des valeurs de similarité entre les documents pour créer des classes de similarité représentant le domaine. Ensuite, à l'aide d'une technique de regroupement et d'une approche fondée sur un graphique, il est possible de construire une hiérarchie pouvant être représentée par une ontologie légère. (Henry et Sengupta, 2007) proposent d'utiliser une ontologie pour extraire des connaissances d'un référentiel XML d'exemplaire des pièces de Shakespeare. L'approche a été implémentée dans une architecture utilisant des langages du Web sémantique, y compris OWL et RuleML, préparant ainsi l'ontologie pour une utilisation dans le partage de données. Dans le même contexte, l'approche proposée par (Aussenac-Gilles et Kamel, 2009) consiste à prendre en compte la structure du document XML qui apporte des connaissances supplémentaires. Dans un premier temps, une caractérisation de la sémantique du balisage XML et de leurs relations est réalisée. Des règles d'analyse

2. sont emboîtés récursivement, décrivant une structure d'arbre. Les éléments possèdent un nom qui est le nom de la balise et peuvent contenir des attributs.

3. possède un nom et une valeur de type atomique (non structurée par des balises, exemple une chaîne de caractère)

4. le format XHTML est l'adaptation du standard HTML aux contraintes syntaxiques du format XML

sont ensuite définies pour exploiter la structure XML des documents et créer des concepts d'ontologie et des relations sémantiques. Dans un deuxième temps, cette ontologie est enrichie à l'aide des résultats de l'analyse de texte par motifs lexico-syntaxiques.

D'autres approches à base de règles d'association existent. Citons l'approche proposée par (Pravin *et al.*, 2014) qui consiste à produire des règles de support et de confiance qui fournissent une connaissance approchée et approximative de la structure et le contenu des documents XML. (Tekli *et al.*, 2016) proposent une approche de désambiguïsation sémantique intitulée XSDF prenant en entrée un document XML, puis produisant en sortie un arbre XML sémantiquement augmenté de termes non ambigus.

Autres travaux. D'autres applications web peuvent servir à structurer et diffuser l'information. Les wikis par exemple sont des moyens établis pour la création collaborative, la gestion des versions et la publication d'articles sous forme de textes. Le projet Wikipédia, par exemple, a réussi à créer l'encyclopédie de loin la plus importante sur la base d'un wiki. Récemment, plusieurs approches ont été proposées sur la façon d'étendre les wikis pour permettre la création de contenu structuré et sémantiquement enrichi. (Auer et J. Lehmann, 2007) proposent une méthode pour expliciter ce contenu structuré par l'extraction d'informations à partir d'instances de modèle de wiki. L'approche consiste à interroger efficacement la grande quantité d'informations extraites (par exemple, plusieurs millions de déclarations RDF pour la version Wikipedia anglaise uniquement), ce qui conduit à des possibilités de réponse aux requêtes. Plus récemment, quelques travaux se sont intéressés à l'exploitation de la structure de wikitext de Wikipédia. (Arnold et Rahm, 2014) proposent une approche utilisant des modèles sémantiques et des techniques NLP (Natural Language Processing) pour traiter les définitions de Wikipedia et identifier les relations sémantiques entre les concepts. L'approche est capable d'extraire plusieurs relations à partir d'un seul article de Wikipedia. (Nugues, 2016) propose d'identifier les entités et les relations entre elles à partir de grands corpus, généralement des versions complètes de Wikipedia composées de millions de documents et du texte. L'approche consiste à appliquer les techniques de traitement du langage naturel pour réaliser une analyse syntaxique et sémantique.

D'autres approches ont été proposées pour répondre au problème de l'augmentation significatif de la quantité de données à stocker et à traiter qui augmente de manière significative (Namici, 2018). L'objectif est d'extraire les connaissances à partir de sources structurées telles que les bases de données relationnelles, les formats CSV, etc. (Unbehauen *et al.*, 2012). Il s'agit le plus souvent d'une conversion des données structurées en RDF et OWL. Citons à titre d'exemple le langage R2RML (actuellement normalisé par le W3C) qui permet de mettre en correspondance le contenu relationnel à RDF (Salas *et al.*, 2011).

2.2.2.2 Approches sur des textes avec mise en page

Dans cette section, nous présentons les approches qui exploitent les **langages de description de page** (*page description languages*) pour l'extraction de connaissances. Ces formats sont soit textuels (par exemple Postscript), soit binaires (par exemple PDF). Cependant, d'autres formats binaires et propriétaires combinant le visuel et l'organisation logique des données sont apparus dans les années 90 telles que le format

DOC de Microsoft.

le format PDF le PDF (*Portable Document Format*) a été introduit en 1993 par la société Adobe avec la version 1.0. Jusqu'à l'apparition de PDF, PostScript (Adobe, 1985) était le format de choix pour partager des articles LATEX avec des diagrammes ou des formulaires complexes. PostScript est le langage de programmation et de description de pages par excellence des plateformes et périphériques, introduit par Adobe 1982. Il décrit le placement et les formes de tous les éléments du document, y compris les polices. Les documents placés sur le Web au format PostScript peuvent être téléchargés sur n'importe quel ordinateur et être imprimés de la même manière sur toutes les imprimantes PostScript. Cependant, l'utilisation de fichiers PostScript sur le web présente plusieurs inconvénients : (i) les fichiers ont tendance à être très volumineux ; (ii) ils ne peuvent pas être visualisés tant que le fichier entier n'a pas été téléchargé ; (iii) si un fichier PostScript n'inclut pas une police particulière utilisée dans le document et que cette police n'est pas installée sur votre ordinateur, une autre police, généralement Courier, est remplacée, ce qui entraîne un rendu inacceptable (Grätzer, 2016).

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, quelle que soit la plateforme ou le programme informatique dans lequel il est visualisé (Merz, 1998). 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. 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 (Ronzano et Saggion, 2016). Les approches mises en place sont personnalisées et sont fondées sur des éléments structurels comme le titre, les sections, les figures, les tableaux, etc. Dans tous les cas, pour automatiser l'extraction d'information, il convient de convertir ces fichiers dans un format exploitable par la machine.

(Klampfl *et al.*, 2014) proposent deux méthodes de reconnaissance de tables fondées sur des techniques d'apprentissage non-supervisées et des heuristiques qui détectent automatiquement l'emplacement et la structure des tables dans un article au format PDF. L'approche est réalisée en deux étapes : (i) identifier à l'aide d'algorithmes les boîtes englobantes des tables individuelles à partir d'un ensemble de blocs de texte étiquetés ; (ii) extraire la structure tabulaire sous forme d'une grille rectangulaire de cellules de tableau à partir de l'ensemble de mots contenus dans ces régions de table.

(Jian *et al.*, 2015) proposent PDFMEF, un cadre d'extraction des connaissances multi-entités pour les documents scientifiques au format PDF. C'est une boîte à outils qui encapsule les outils d'extraction open source. Par exemple, il exploite PDFBox et TET pour l'extraction de texte intégral, GROBID pour l'extraction d'en-têtes, etc.

(Ahmad *et al.*, 2016) proposent une approche pour pouvoir extraire des informations dans des articles de recherche publiés au format PDF. L'approche consiste : (i) convertir les documents vers le format XML ; (ii) exploiter les formats XML et texte à l'aide de règles et d'heuristiques pour extraire les informations.

2.2.3 Discussion et positionnement

Dans les sections précédentes, nous avons situé le contexte de ce travail, qui est relatif à la présentation des résultats en statistiques. Nous avons identifié deux problématiques à considéré : (i) interprétation des informations contenues dans les tableaux selon le processus humain d'interprétation des tableaux statistiques (voir figure 2.1) ; (ii) automatisation de ce processus d'interprétation à partir de documents au format PDF, qui nécessite l'identification et l'extraction des tableaux pertinents. Nous avons présenté deux familles d'approches pour l'extraction des connaissances à partir de textes : les approches utilisant le contenu textuel non-structuré (sans mise en forme et sans mise en page) et les approches exploitant le contenu textuel structuré (à l'aide d'une mise en forme ou une mise en page).

Le but de ce travail est d'extraire automatiquement des connaissances à partir de textes. Dans notre étude, nous nous intéressons plus précisément à l'extraction et l'interprétation de connaissances à partir de tableaux statistiques publiés au format PDF. Le fait de travailler essentiellement à partir de tableaux statistiques se justifie : (i) le cadre d'évaluation du risque cardiovasculaire qui est souvent celui de l'épidémiologie ; (ii) les tableaux sont souvent utilisés pour présenter le résultat des études statistiques. Les connaissances sont en général implicites dans les tableaux et notre objectif est d'extraire ces connaissances pour pouvoir les exploiter. Par ailleurs, une des caractéristiques des tableaux statistiques est une représentation synthétique et visuelle de résultats statistique contenant des informations liées. Le problème est l'extraction des informations pertinentes ainsi que le lien entre ces dernières. Une deuxième difficulté, évoquée dans la section précédente, vient du fait que les études statistiques sont souvent publiées sous forme de documents. Ce constat nous a conduits à envisager une approche permettant de convertir le PDF vers un format exploitable toute en préservant la structure des tableaux.

2.3 Approche proposée

La méthodologie (voir figure 2.5) consiste à extraire les connaissances à partir de présentation statistique de résultats : (i) convertir cette dernière vers un format exploitable ; (ii) localiser et extraire la connaissance à l'aide d'un vocabulaire contrôlé et des techniques de traitement automatique de la langue (TAL) ; (ii) expliciter les liens sous forme de triplets RDF. Cette approche présente l'avantage de prendre en charge des modèles statistiques sous différentes mises en forme en entrée de traitement.

Dans cette approche, la première tâche concerne les documents de présentation de résultats statistiques publiés au format PDF et DOC. Elle consiste à convertir ces documents vers un format exploitable. Les tâches suivantes : la localisation d'information, l'extraction d'information et la construction des triplets sont communes aux différents formats. Elles utilisent un vocabulaire contrôlé et des techniques de traitement automatique de la langue.

2.3.1 Conversion des documents de présentation de résultats statistique

Notre première tâche a été de convertir les documents pour pouvoir exploiter le contenu des tableaux par des traitements automatiques. La littérature fait état de nombreux outils et de méthodes de conversion de format. Nous centrons notre propos

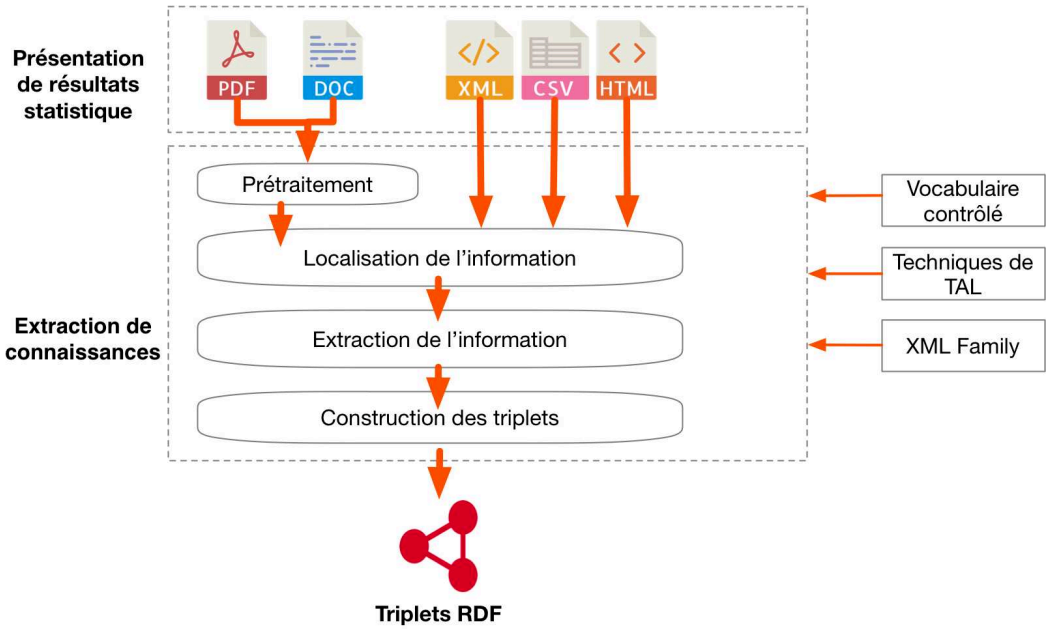


FIGURE 2.5 – Etapes de la méthodologie proposée

sur les outils prenant en entrée des documents au format PDF. Dans ce contexte, pour justifier le choix de l'outil que nous avons utilisé, nous proposons une analyse de ceux souvent cités comme référence :

- pdftohtmlEX [5] convertit les fichiers PDF au format HTML en conservant le texte et la mise en forme des tableaux ;
- pdftohtml [6] convertit les fichiers PDF au format HTML et XML ;
- PDFX [7] utilise des règles pour reconstruire la structure logique d'articles scientifiques au format PDF, quel que soit leur style de formatage ;
- Tabula [8] extrait semi-automatiquement des tableaux de données à partir de fichiers PDF ;
- PDFMiner [9] extrait des informations à partir de documents PDF. Contrairement à d'autres outils liés au PDF, il est entièrement consacré à l'obtention et l'analyse de données de texte ;
- PDF Online [10] convertit les fichiers PDF au format HTML en conservant le texte et la mise en forme des tableaux.
- GROBID [11] permet d'extraire, d'analyser et de restructurer des documents bruts au format PDF en des documents structurés codés au format *Text Encoding Initiative (TEI)*, en mettant l'accent sur les publications techniques et scientifiques.

Une synthèse des caractéristiques de ces outils est présentée dans le tableau de la figure 2.1. Nous avons sélectionné l'outil pdftohtmlEX (voir figure 2.6-(A)), pour les raisons suivantes : (1) le format de sortie HTML préserve la structure tabulaire lors du

5. http://coolwanglu.github.io/pdf2htmlEX/
6. http://pdftohtml.sourceforge.net/
7. http://pdfx.cs.man.ac.uk/
8. http://tabula.technology/
9. http://www.unixuser.org/~euske/python/pdfminer/
10. http://www.pdfonline.com/
11. https://grobid.readthedocs.io/en/latest/

	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	GPLv2	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 BCL technologie	-	Oui	HTML	-	Non	automatique
GROBID	Service web, Api	Apache 2.0	Python, Java, Node.js	Oui	XML	16/08/2018	Expérimental	automatique

TABLE 2.1 – Tableau comparatif des outils de conversion de documents PDF

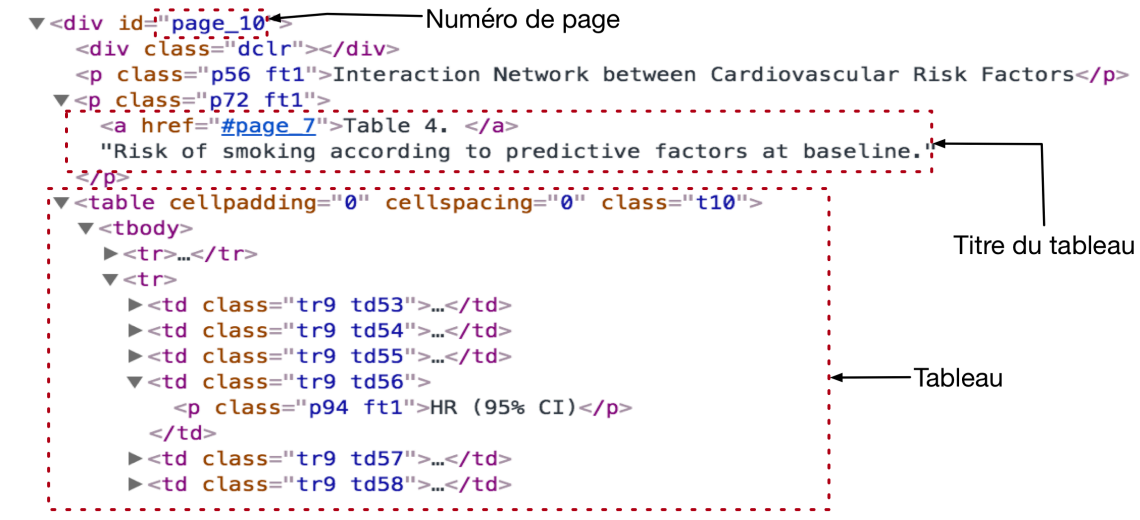
processus de conversion contrairement à PDFX (voir figure 2.6-(B)) ; (2) l'information concernant la position du tableau dans le document est disponible. Ainsi, il est possible d'effectuer l'action inverse en partant des éléments extraits il est possible d'identifier l'endroit du tableau contenant cette information ; (3) la conversion est automatique contrairement à l'outil Tablula. En revanche, la sortie HTML obtenue présente l'inconvénient d'être linéaire et nécessite des traitements supplémentaires.

2.3.2 Localisation et extraction de l'information

2.3.2.1 Localisation de l'information.

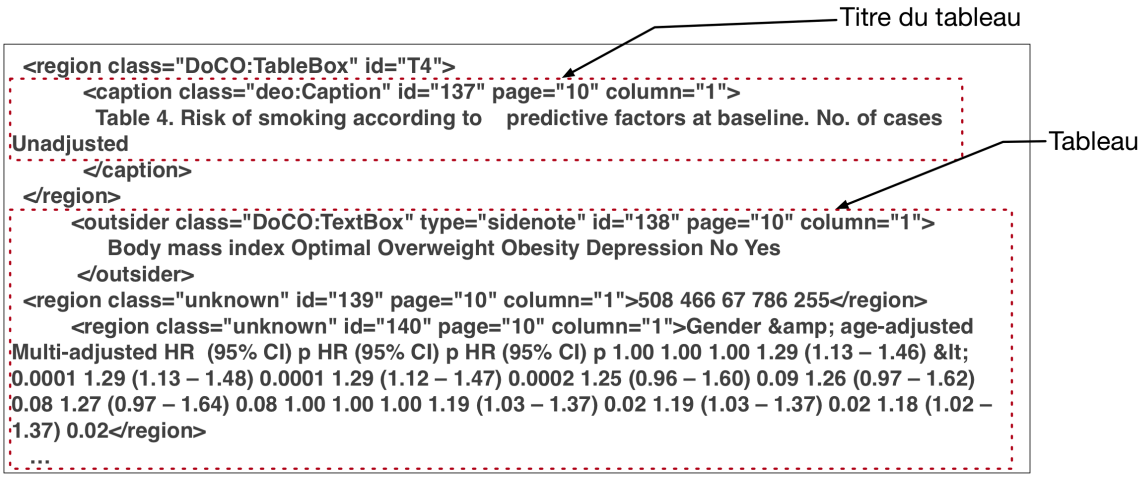
Une des difficultés de ce travail concerne la reconnaissance de tableaux contenant des informations relatives à un contexte d'étude. Au cours de l'analyse visuelle d'un tableau, il est relativement aisé pour un humain d'identifier facilement le sujet étudié. Ce processus de déduction est réalisé en identifiant certains termes dans le titre, la 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) [12] ou de trouver des heuristiques comme l'ont suggéré (SHIGAROV, 2015) et (CLARK et DIVVALA, 2015), nous avons opté pour une approche supervisée exploitant un vocabulaire contrôlé du domaine. 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 ou le corps du tableau. Contrairement aux deux approches qui supposent un travail important d'élaboration d'heuristiques et de règles, notre démarche est générique et ne nécessite qu'un prétraitement léger. Par exemple, il suffit de convertir le document sans étudier la disposition de chaque type de tableaux présent dans le document pour construire des heuristiques adaptées à chacun.

12. propose une approche formelle pour mapper des données tabulaires au format RDF. Les mappages par défaut sont automatiquement créés et peuvent être révisés par la communauté en utilisant un wiki sémantique. Les mappages sont exécutés à l'aide d'une conversion RDB2RDF.




```
<div id="page_10">
  <div class="dclr"></div>
  <p class="p56 ft1">Interaction Network between Cardiovascular Risk Factors</p>
  <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">
    <tbody>
      <tr>…</tr>
      <tr>
        <td class="tr9 td53">…</td>
        <td class="tr9 td54">…</td>
        <td class="tr9 td55">…</td>
        <td class="tr9 td56">
          <p class="p94 ft1">HR (95% CI)</p>
        </td>
        <td class="tr9 td57">…</td>
        <td class="tr9 td58">…</td>
```

(A) Format de sortie Pdf2htmlEX




```
<region class="DoCO:TableBox" id="T4">
    <caption class="deo:Caption" id="137" page="10" column="1">
      Table 4. Risk of smoking according to  predictive factors at baseline. No. of cases
Unadjusted
    </caption>
</region>
    <outsider class="DoCO:TextBox" type="sidenote" id="138" page="10" column="1">
      Body mass index Optimal Overweight Obesity Depression No Yes
    </outsider>
<region class="unknown" id="139" page="10" column="1">508 466 67 786 255</region>
    <region class="unknown" id="140" page="10" column="1">Gender &amp;amp; age-adjusted
Multi-adjusted HR  (95% CI) p HR (95% CI) p 1.00 1.00 1.00 1.29 (1.13 – 1.46) &amp;lt;
0.0001 1.29 (1.13 – 1.48) 0.0001 1.29 (1.12 – 1.47) 0.0002 1.25 (0.96 – 1.60) 0.09 1.26 (0.97 – 1.62)
0.08 1.27 (0.97 – 1.64) 0.08 1.00 1.00 1.00 1.19 (1.03 – 1.37) 0.02 1.19 (1.03 – 1.37) 0.02 1.18 (1.02 –
1.37) 0.02</region>
  ...
```

(B) Format de sortie pdfx en xml

FIGURE 2.6 – Sortie obtenue avec PdftohtmlEX et PDFX

2.3.2.2 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 de les structurer pour expliciter les liens entre ces informations. L'extraction des informations à partir d'un tableau nécessite deux étapes : (i) la détermination des colonnes pertinentes qui est obtenue en utilisant un vocabulaire contrôlé. Ainsi, au lieu d'extraire l'ensemble des colonnes et se retrouver avec un nombre important de triplets à traiter, nous ciblons les colonnes à extraire. L'avantage de cette approche est quelle permet d'accéder à toutes les informations contenues dans les tableaux, pour cela il suffit d'ajouter le terme au vocabulaire contrôlé et refaire l'extraction ; (ii) 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 est constituée de quatre étapes :

1. **L'extraction des tableaux pertinents** comporte deux phases : (a) la reconnaissance des tableaux ; (b) la vérification de la pertinence du tableau. La recon-

naissance 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. La pertinence est déterminée par le nombre de tags réalisé sur le tableau, ainsi l'utilisateur peut fixer une valeur à partir de laquelle un tableau est considéré comme pertinent. Par exemple, si le nombre de tags est supérieur à quatre alors le tableau est pertinent.

2. **L'extraction des colonnes pertinentes** consiste à extraire les colonnes du tableau 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 de 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 ».

2.3.3 Construction des triplets

À l'issue des deux étapes précédentes, on obtient une collection de tableaux nettoyés et exploitables. L'objectif de cette étape est de décrire l'ensemble des tableaux pertinents à l'aide de triplets appelé graphe RDF. L'approche choisie est décrite dans la

figure 2.7 (A). L'automatisation du processus de construction des paraphrases (décrite dans la section 1.1.3) est réalisée de la manière suivante :

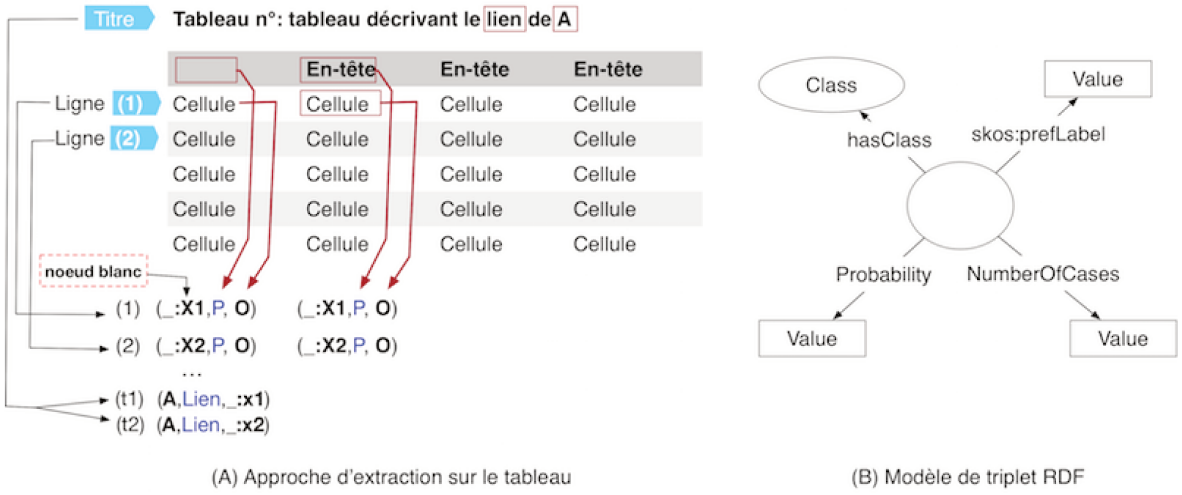


FIGURE 2.7 – Stratégie d'extraction des triplets à partir des tableaux

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 à 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 répandues 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 triplets guidant leur extraction (voir figure 2.7 (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 d'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 [13]. A partir de l'élément et de la relation identifiés dans le titre F, des triplets sont construits selon le modèle (s_t, p_t, o_t) où :

- s_t correspond à l'élément identifié dans le titre ;
- p_t correspond au type de relation entre le titre et le tableau ;
- o_t correspond au nœud blanc « _ :x_i » pour chaque ligne du tableau T.

13. `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

2.4 Expérimentation

La première expérimentation vise à évaluer l'approche d'extraction des connaissances élaborée et détaillée dans la section précédente. L'objectif est d'estimer d'une part les performances d'extraction des informations à partir des tableaux pertinents dans des documents au format PDF et, d'autre part, la généricité de l'approche.

Pour réaliser ces objectifs, nous proposons d'extraire automatiquement les informations contenues dans des tableaux statistiques sous forme de triplets RDF pour notre domaine d'application (domaine des maladies cardiovasculaires) et pour un autre domaine (domaine financier). Nous nous focalisons sur la reconnaissance des tableaux pertinents et leur extraction sous forme de triplets.

2.4.1 Domaine des maladies cardiovasculaires

Pour conduire notre expérimentation, nous travaillons à partir d'une étude statistique menée par (Meneton *et al.*, 2016) dans le domaine de l'épidémiologie. L'objectif de l'étude 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. L'étude est publiée sous forme d'un document PDF dans lequel 13 tableaux décrivent les associations entre 13 facteurs de risque cardiovasculaires. La première étape a consisté à convertir le fichier PDF en 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 du document PDF (voir figure 2.8). L'ensemble des tableaux décrivant les associations entre les facteurs de risque cardiovasculaires ont tous la même structure.

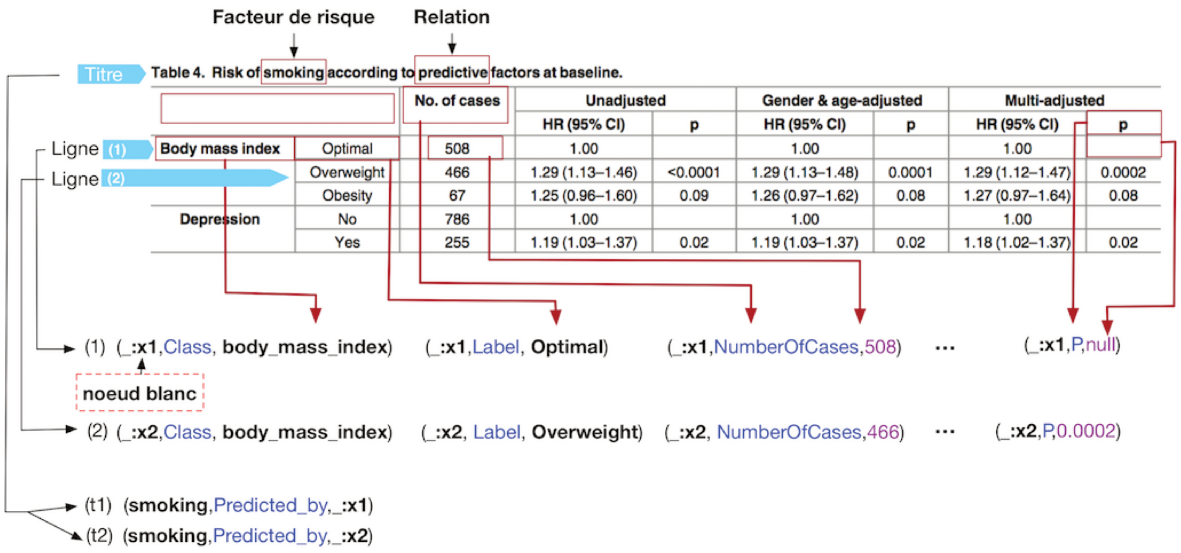


Figure 2.8 – Stratégie d'extraction de l'information appliquées sur le tableau

Chaque ligne i du tableau comporte un titre (voir figure 2.8). 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 :

— « _ :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.

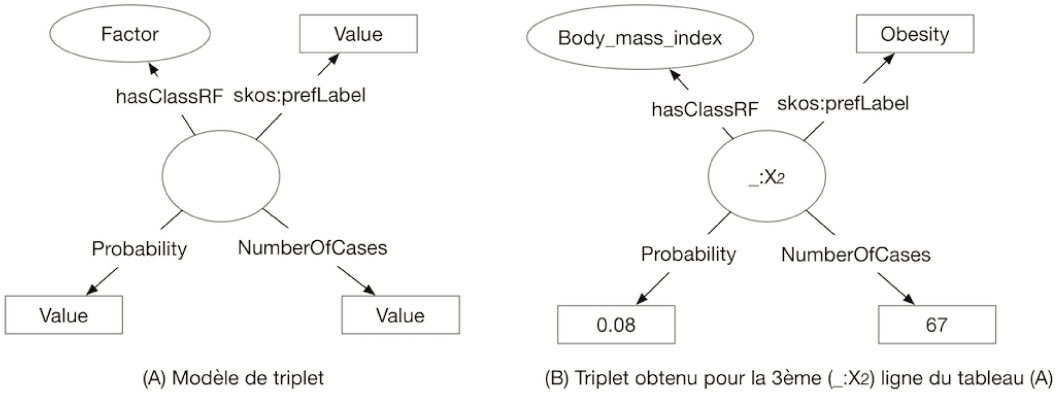


FIGURE 2.9 – Modèle de triplets pour l'extraction d'information

Chaque ligne du tableau est décrite par un triplet suivant le modèle décrit dans partie (A) de la figure 2.9. Par exemple, le résultat obtenu pour la troisième ligne du tableau est présenté dans la partie (B) de la figure 2.9.

— **Construction du vocabulaire contrôlé** : L'analyse de la littérature concernant les vocabulaires décrivant les facteurs de risque cardiovasculaires nous a permis de constater que ces derniers ne décrivent que d'une manière *partielle* les interactions entre les facteurs de risque cardiovasculaires. « CARRE [14] » 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 (voir figure 2.10) en utilisant les termes du domaine des maladies cardiovasculaires utilisés par les experts du domaine. L'objectif est d'exploiter ce vocabulaire pour identifier, de façon précise, les données apparaissant dans un tableau.

  Pour construire notre vocabulaire contrôlé, nous avons adopté la démarche suivante : (i) identifier les termes (mots-clés) le plus souvent utilisés dans la littérature pour décrire les facteurs de risque cardiovasculaire ; (ii) identifier la structure qui rendra les termes utilisables. Une liste de termes est un début, mais ils ne sont probablement pas organisés dans un système qui les rend faciles à mémoriser, à utiliser et à mettre à jour. Un vocabulaire contrôlé bien conçu nécessite une structure regroupant des termes pertinents dans une hiérarchie. Cette structure facilitera la recherche des termes et nous aidera à voir quand et où nous avons besoin de plus (ou de moins) de termes. Une fois le vocabulaire contrôlé établi, nous pouvons maintenir la terminologie à jour, ajouter des termes lorsque le besoin se fait sentir, et également supprimer ou modifier des termes lorsqu'ils ne sont plus utilisés dans le domaine.

14. `http://bioportal.bioontology.org/ontologies/CARRE`

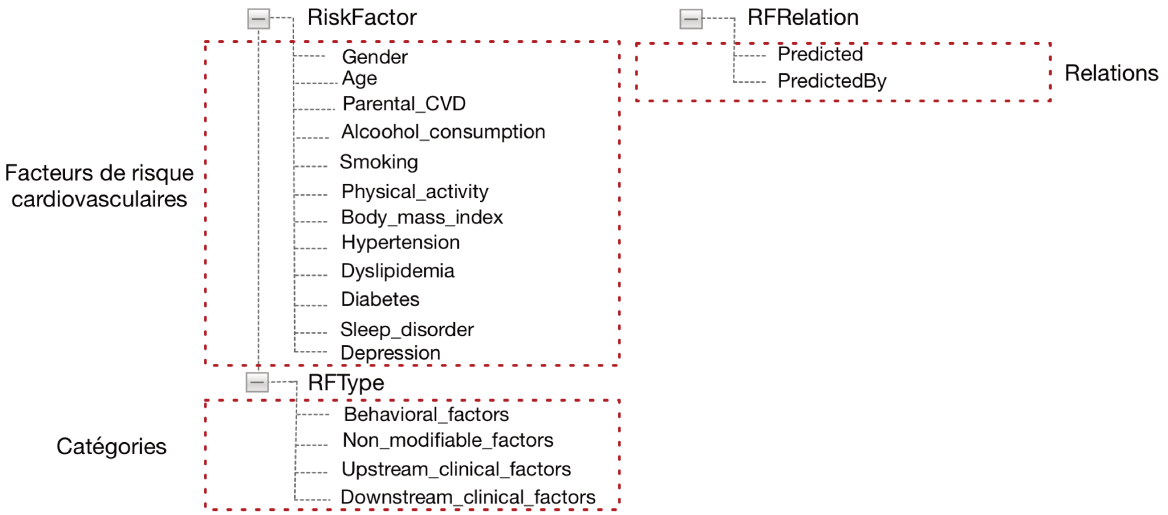


FIGURE 2.10 – Vocabulaire contrôlé construit

Les termes figurant dans le vocabulaire contrôlé appartiennent à l'une des deux catégories suivantes :

- RiskFactor, cette catégorie liste les facteurs de risque cardiovasculaire le plus souvent étudiés dans la littérature. Cependant, cette liste n'est pas exhaustive ;
- RFType, cette catégorie liste les différentes catégories associées au facteur de risque cardiovasculaire ;
- RFRelation, cette catégorie liste les différentes relations entre les facteurs de risque. Pour le moment, nous n'avons identifié que deux relations. Cette catégorie sera enrichie par la suite.

— **Localisation de l'information** : le résultat obtenu à partir du tableau (A) est présenté dans la figure 2.11. 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 2.11 – Exemple de tags 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 « *numero_de_page* => *titre* => *contenu* » (voir figure 2.12).

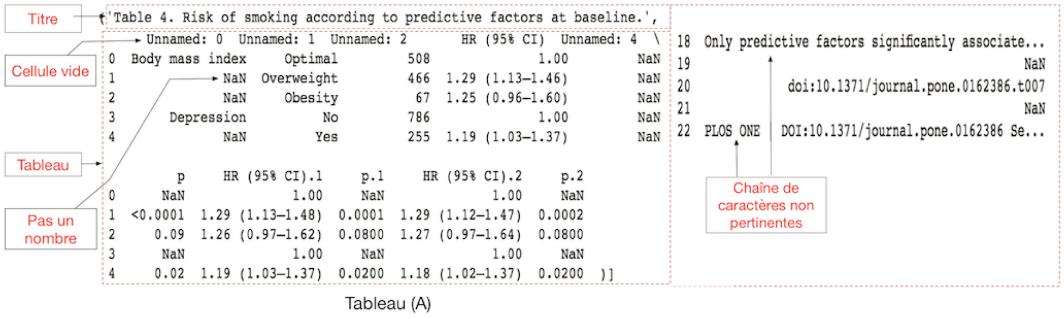


FIGURE 2.12 – Exemple de tableau extrait à partir du document HTML

— **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 2.13 – Tableau (A) de la figure 2.12 après nettoyage

— **Nettoyage et mise en forme du tableau** : les résultats obtenus pour le tableau (A) sont présentés dans la figure 2.13. 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 2.14. Nous avons obtenu trois nœuds blancs ($_ :x_1, _ :x_2, _ :x_3$). 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
($'Smoking', 'PredictedBy', _ :x_1$).

2.4.2 Domaine financier

Pour conduire notre expérimentation, nous travaillons à partir d'une enquête internationale sur les transactions de change et de produits dérivés [15] dans le domaine

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

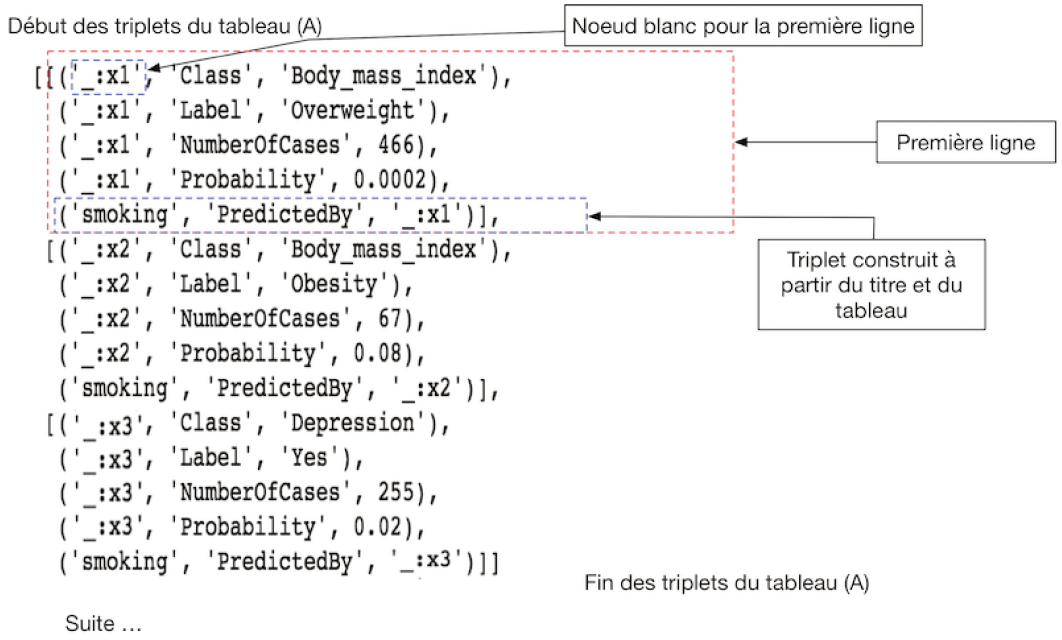


FIGURE 2.14 – Triplets obtenus à partir du tableau (A) de la figure 2.12

financier. Nous avons appliqué le même traitement que lors de l'expérimentation précédente. La première étape a consisté à convertir le fichier PDF en un format exploitable à l'aide de pdftohtmlEX. Le résultat obtenu est un document HTML contenant « 2330 lignes », « 21774 mots », « 298187 caractères » et « 19 tableaux ». Ensuite, les tableaux pertinents ont été extraits sous forme de triplets. La figure 2.15 présente l'exemple d'un tableau extrait du document. Le résultat obtenu après extraction des triplets est présenté dans la figure 2.16.

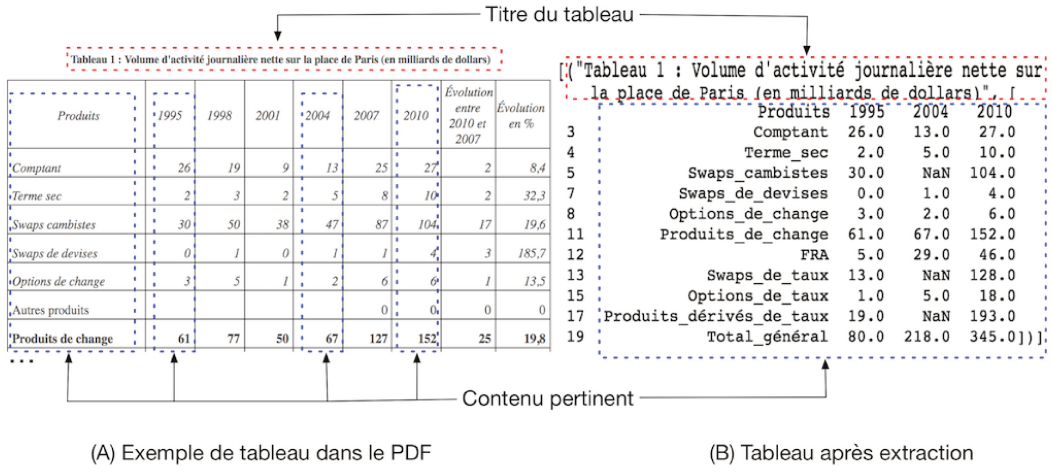


FIGURE 2.15 – Extraction d'un tableau figurant dans l'enquête

2.5 Evaluation de l'approche

Nous avons montré dans la section précédente comment extraire et expliciter 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

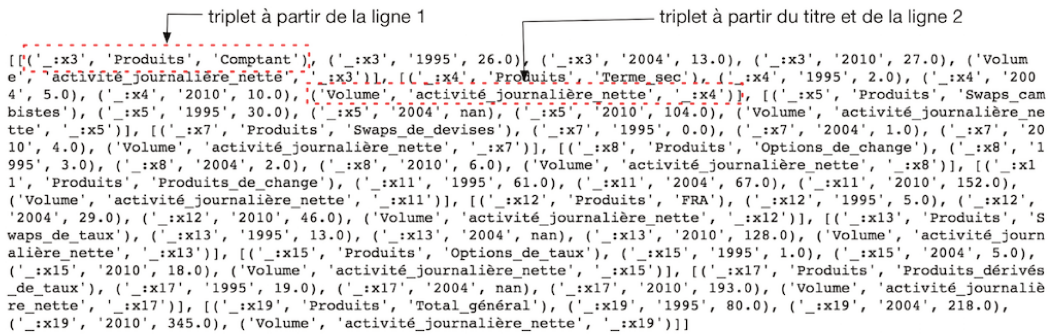


FIGURE 2.16 – Liste des triplets obtenus à partir du tableau de la figure 2.15(B)

étapes : (i) localisation de l'information dans les tableaux ; (ii) extraction de l'information sous forme de triplets RDF. Ces deux étapes sont fondées sur l'utilisation d'un vocabulaire contrôlé. Pour permettre une première évaluation, nous avons conduit deux évaluations : la première concerne l'extraction des tableaux et la seconde concerne la pertinence des triplets extraits.

Evaluation de l'extraction. Pour évaluer l'approche d'extraction des tableaux, nous avons conduit une évaluation fondée sur l'interprétation par un expert de la sélection de tableaux statistiques correspondant à un sujet d'étude dans un document PDF. Pour cela, nous avons travaillé sur deux jeux de données disponibles au format PDF : le premier (D1) concerne le modèle statistique des facteurs de risque dans le domaine des maladies cardiovasculaires, la démarche et les résultats obtenus sont présentés dans la section 2.6.1 ; le second (D2) concerne l'enquête internationale sur les transactions de change et de produits dérivés présentés en section 2.6.2.

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 identifiés comme traitant des interactions entre les facteurs de risque cardiovasculaire (D1) et T_{vat} le nombre de tableaux traitant de volume d'activité de transactions (D2).

La précision de l'extraction pour le jeux de donnée (D1) est calculé selon :

$$Précision = \frac{T_{fac}}{\text{nombre total de tableaux apparaissant dans (D1)}} \tag{2.1}$$

La précision de l'extraction pour le jeux de donnée (D2) est calculé selon :

$$Précision = \frac{T_{vat}}{\text{nombre total de tableaux apparaissant dans (D2)}} \tag{2.2}$$

Le rappel de l'extraction pour le jeux de donnée (D1) est calculé selon :

$$Rappel = \frac{T_{fac}}{\text{nombre de tableaux décrivant réellement les interactions entre FRC}} \tag{2.3}$$

Le rappel de l'extraction pour le jeux de donnée (D2) est calculé selon :

$$Rappel = \frac{T_{vat}}{\text{nombre de tableaux décrivant réellement le volume d'activité de transactions}} \tag{2.4}$$

La F-mesure pour (D1) et (D2) est calculé selon :

$$F\text{-}mesure = \frac{2 * Précision * Rappel}{Rappel + Précision} \tag{2.5}$$

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 2.2 présente les résultats de l'extraction des tableaux avec

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 (10 termes)	**0.76**	**0.69**	**0.72**
	avec (20 termes)	**0.89**	**0.80**	**0.84**

TABLE 2.2 – Résultats expérimentaux

les mesures de *précision*, *rappel* et *F-mesure* pour les jeux de données (D1) et (D2). Nous constatons que la précision 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 jeu de données (D1) fournit une performance presque parfaite. 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. Par ailleurs, nous avons été assez surpris par les résultats très satisfaisants de la méthode sur le jeu de données (D2), nous l'expliquons par une sélection restrictive et précise des motifs recherchés à l'aide du vocabulaire. 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 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.

Evaluation de la pertinence des triplets extraits. Pour évaluer la pertinence des triplets extraits, nous avons réalisé l'extraction sur un second jeu de données (D3) dans le domaine des maladies cardiovasculaires. Ce jeu de données est un ensemble de tableaux statistiques décrivant des interactions entre les facteurs de risque cardiovasculaires. Ces résultats ont été obtenus par le LIMICS en collaboration avec le premier auteur de l'étude (D1) à partir de l'étude SUVIMAX [16] réalisée par l'EREN [17] dans le domaine de l'épidémiologie nutritionnelle. L'étude réalisée par le LIMICS a consisté à examiner toutes les relations existantes entre les principaux facteurs de risque de maladies cardiovasculaires dans cette cohorte. Les tests de régression de Cox ont permis de vérifier systématiquement les associations prospectives entre facteurs de risque comportementaux et cliniques (sexe, âge, antécédents parentaux de maladie cardiovasculaire, consommation d'alcool non modérée, tabagisme, inactivité physique, obésité, hypertension, dyslipidémie, diabète, trouble du sommeil, dépression, alimentation malsaine).

Le résultat obtenu à partir de (D3) présente des variations et des similitudes par

16. `https://fr.wikipedia.org/wiki/SU.VI.MAX`
17. `https://eren.univ-paris13.fr/index.php/fr/`

rapport au modèle de facteurs de risques obtenus à partir de (D1). Par exemple, dans la figure 2.17 le facteur de risque « Smoking » est prédit par « Depression » et « Obesity » dans (D1) et prédit par « Age », « Gender », « AlcoholConsumption » dans (D3). Un second exemple est présenté dans la figure 2.18, celui-ci montre que le facteur « Obesity » est prédit par les facteurs « Dyslipidemia », « PhysicalInactivity », « Smoking » dans (D1) et (D3). En revanche, une variation est observée pour les facteurs « Depression » et « Hypertension » dans (D1), et « Age », « Gender », « AlcoholConsumption » dans (D3).

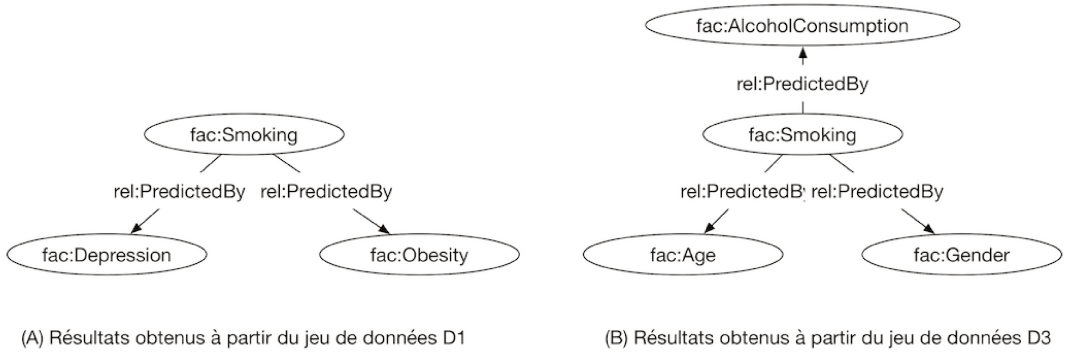


FIGURE 2.17 – Résultats obtenus pour le facteur de risque « Fumeur » pour les jeux de données (D1) et (D3)

Trois explications permettent de comprendre ces variations :

— similarité, concerne les facteurs et les relations qui sont déjà validés par des preuves scientifiques ;

— différence, concerne les facteurs et les relations pour lesquels le recours à des experts ou la littérature est nécessaire pour les valider ;

— absence, correspondent a des trous de connaissances, ils concernent les nouveaux facteurs et liens qui ne sont pas encore répertoriés dans la littérature.

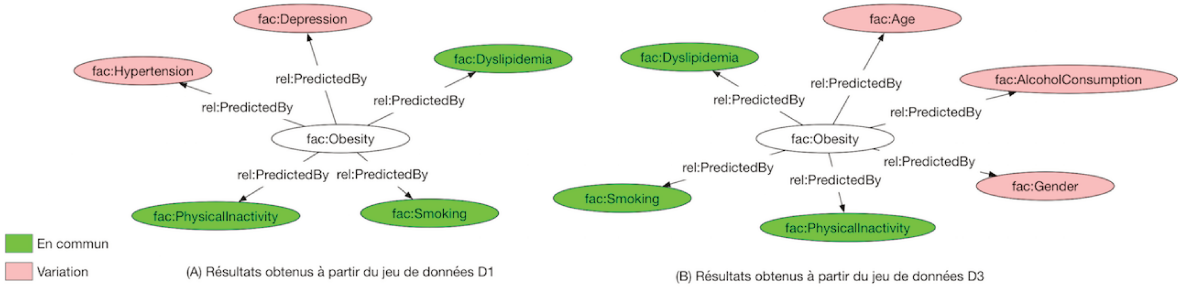


FIGURE 2.18 – Résultats obtenus pour le facteur de risque « Obesity » pour les jeux de données (D1) et (D3)

Ce bilan est essentiel pour la modélisation de connaissances. Il permet à partir des différents modèles obtenus de mettre en évidence le niveau où se situent les connaissances communes qui peuvent représenter le cœur d'un modèle générique. Par exemple, dans la figure 2.18 qui décrit le facteur de risque « Obesity », le cœur commun entre ces deux modèles correspond au facteur de risque « Dyslipidemia, Smoking et PhysicalInactivity ». Ils font également apparaître les concepts absents dans les différents modèles. Ces concepts sont, soit des concepts connus dans la littérature et n'ont pas été pris en compte dans les modèles, soit des trous de connaissances scientifiques. La

définition de ces concepts correspondant à ces trous de connaissance est essentielle pour déterminer des pistes de recherche. Elle est d'autant plus fondamentale que ces travaux ont pour objectifs d'apporter les connaissances essentielles aux processus d'évaluation du risque cardiovasculaire.

2.6 Conclusion

Dans ce chapitre, nous avons décrit une méthode permettant d'extraire des connaissances à partir d'un modèle statistique présenté sous forme de tableau et publié au format PDF, sous forme de triplets RDF. 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.).

En outre, l'approche est actuellement utilisée sur des études statistiques dans 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. Dans le chapitre suivant, nous présenterons l'approche adoptée afin de construire ce modèle générique.

L'apport de l'approche développée dans ce chapitre pour la communauté de l'évaluation du risque cardiovasculaire est important. Elle permet de représenter les interactions entre les facteurs de risque cardiovasculaire dans un modèle exploitable dans des systèmes d'évaluation du risque. A notre connaissance, aucun modèle de ce type n'existe encore dans l'évaluation du risque cardiovasculaire.

Chapitre 3

Modélisation des connaissances : application au facteurs de risque cardiovasculaire

Sommaire

3.1 Définition et objectifs **66**
3.1.1 La notion de connaissance 66
3.1.2 Ingénierie des connaissances 66
3.1.3 La modèlisation des connaissances 67
3.1.4 Objectif 68
3.2 Approche proposée **69**
3.2.1 E1 : construction du Premier modèle conceptuel 70
3.2.2 E2 : définition du modèle conceptuel complet 71
3.2.3 Discussion 74
3.2.4 E3 : définition et validation du modèle conceptuel complet .. 74
3.3 Conclusion **76**

Dans ce chapitre, nous présentons un modèle conceptuel générique décrivant les facteurs de risque et leurs interactions. Nous abordons la problématique de modélisation conceptuelle des connaissances en nous fondant sur les concepts de l'ingénierie des connaissances afin d'offrir une modélisation consensuelle et partagée sur les maladies cardiovasculaires. Nous détaillons et discutons la démarche adoptée afin de démontrer l'utilité d'un modèle conceptuel générique pour un domaine.

Textes de références :

Rabia Azzi, Sylvie Despres, Jérôme Nobécourt. D'un modèle statistique à un modèle de connaissance : retour d'expérience. Sylvie Ranwez. 29es Journées Francophones d'Ingénierie des Connaissances, IC 2018, pp.105-119, 2018 Nancy, France.

3.1 Définition et objectifs

3.1.1 La notion de connaissance

A la différence de l'information qui a été définie par (Shannon et Weaver, 1949) en 1949, et constitue le fondement de ce qu'on appelle la théorie de l'information, il n'existe visiblement pas de définition scientifique de la connaissance. Plusieurs définitions dans la littérature existent, qui dépendent essentiellement du contexte d'utilisation. Nous n'avons pas l'intention ici d'étudier ces définitions ni d'en proposer une nouvelle. Par contre, nous allons essayer de caractériser cette connaissance, par rapport à notre approche et nos travaux de thèse.

On se place dans le contexte particulier, proposé par (Ermine *et al.*, 1996), dans lequel la connaissance peut être perçue selon l'hypothèse *sémiotique*, à savoir le triangle sémiotique et la conjonction S3 : signe, signifié, et signifiant (voir figure 3.1). Il s'agit de percevoir la connaissance comme un signe (voir figure 3.2) qui contient de l'*information* (quelle est la forme codée ou perçue du signe que je reçois ?), du *sens* (quelle représentation l'information engendre-t-elle dans mon esprit ?), et du *contexte* (quel environnement conditionne le sens que je mets sur l'information reçue ?) (Aries *et al.*, 2008). On distingue dans cette conception, que la connaissance ne se résume pas à l'information, mais plutôt à la conjonction de trois points de vue extrêmement liés.

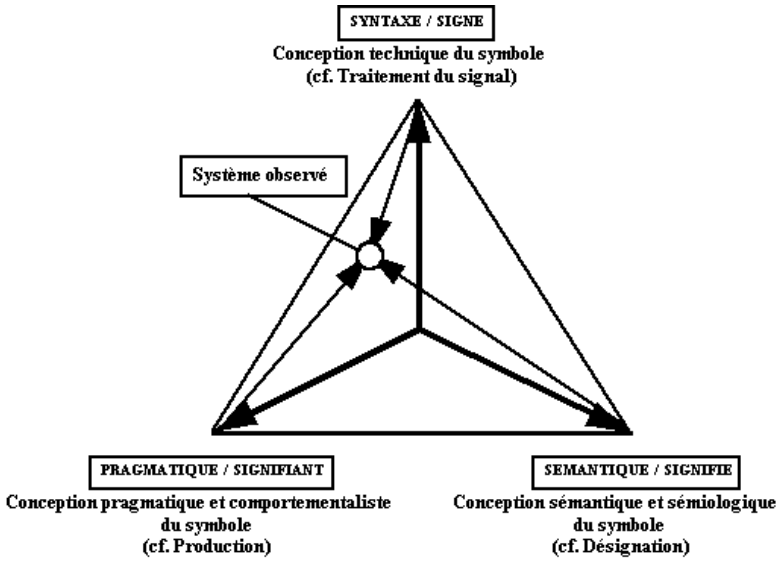


Figure 3.1 – Les trois axes du triangle sémiotique, la conjonction S3 (Le Moigne, 1990)

En effet dans les modèles que nous allons présenter dans la suite de ce chapitre, nous allons manipuler des éléments de la connaissance (concepts, entités, relations, etc.) avec une syntaxe, une sémantique et un contexte d'utilisation bien définis.

3.1.2 Ingénierie des connaissances

L'ingénierie des connaissances (IC) a émergé fin des années 80 lorsqu'à la suite de travaux en Intelligence Artificielle (IA), on s'est posé le problème de formaliser des connaissances (en particulier celles des experts), afin de pouvoir réaliser des systèmes informatisés (Madeleine et Pietri, 2000). Elle contient plusieurs champs de

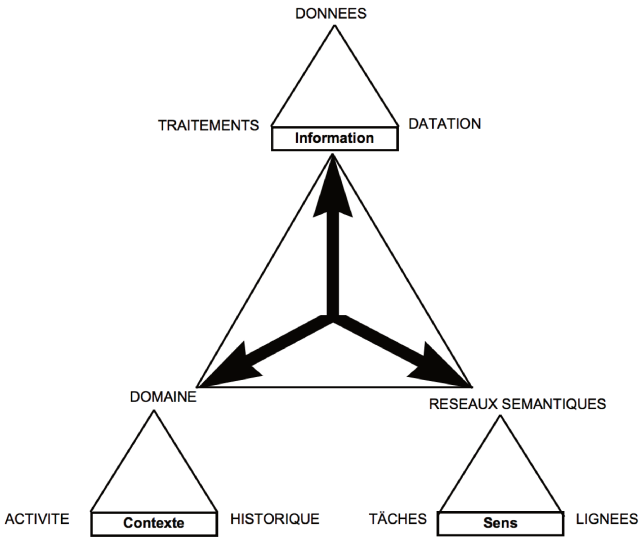


FIGURE 3.2 – Les trois axes du triangle sémiotique selon (ARIES *et al.*, 2008)

recherche en système d'information, modélisation objet, travail collaboratif, interface homme-machine, raisonnement à partir de cas, systèmes à base de connaissances, etc.

Depuis le début des années 2000, la perspective s'est élargie et l'Ingénierie des Connaissances (IC) propose des concepts, méthodes et techniques permettant de modéliser, de formaliser et d'acquérir des connaissances dans les organisations dans un but d'opérationnalisation, de structuration ou de gestion au sens large (AUSSENAC-GILLES, CHARLET *et al.*, 2012).

L'activité de l'IC vise à proposer des concepts, des méthodes et des techniques permettant de **modéliser**, de formaliser, d'**acquérir** des connaissances dans les organisations dans un but d'**opérationnalisation**, de structuration ou de gestion au sens large (par exemple, **valider** les modèles qualitatifs conçus lors de l'étape précédente en conservant une traçabilité entre les modèles au niveau des connaissances et les modèles pour l'implémentation.) (TEULIER *et al.*, 2005). Lors du passage de l'opérationnalisation à la validation, le processus de validation doit être conçu pour prendre en compte les retours des experts et des utilisateurs afin d'améliorer les modèles qualitatifs élaborés et de progresser dans l'analyse du problème et la définition des méthodes élaborées pour le résoudre (CHARLET *et al.*, 2000).

Pour résumer, trois activités principales peuvent être distinguées en IC : la modélisation et l'acquisition des connaissances, l'opérationnalisation des modèles de connaissances, et la validation des méthodes, modèles, outils conçus et opérationnalisés.

3.1.3 La modèlisation des connaissances

La modélisation des connaissances est une activité qui vise à « identifier et structurer les connaissances d'un domaine en une représentation schématique pour les rendre visibles, manipulables, compréhensibles, communicables » (PAQUETTE, 2002).

La modélisation des connaissances peut être définie comme : « un processus de construction intentionnelle (afin de résoudre des problèmes), à partir d'expérience (le savoir-faire), information et savoir, de modèles susceptibles de rendre intelligible (ayant un sens dans un contexte donné) cette connaissance » (BEN-AHMED, 2005).

Il existe un grand nombre de méthodes et de techniques pour la modélisation des connaissances dans la littérature par exemple : (i) la méthode KOD (*Knowledge*

Oriented Design) établie par (Vogel, 1989), elle consiste à développer des systèmes à bases des connaissances ; (ii) la méthode KADS (*Knowledge Analysis and Design System/Support*) conçue en 1985 dans le contexte d'un projet appelé ESPRIT composé de plusieurs experts. Elle a évolué et s'est enrichie dans les années 90. Elle est devenue CommonKADS et cherche à fonder un standard européen (Schreiber *et al.*, 2000). Elle permet la construction de plusieurs modèles (modèle d'organisation, modèle des tâche, modèle conceptuel, etc.) liés entre eux, nécessaires à la spécification d'un système à base de connaissances, dont le modèle organisationnel permet de prendre en compte l'utilisation des connaissances dans leur environnement ; (iii) la méthode MKSM (*Methodology for Knowledge System Management*) conçu par (Ermine *et al.*, 1996) afin de gérer les connaissances des chercheurs. Elle a ensuite été enrichie pour devenir MASK (*Method for Analysis and Structuring Knowledge*). Elle est utilisée dans de nombreuses entreprises en France et dans le monde telles que EDF, Thomson, La Poste, Cofinoga, Gaz de France, etc.

Le modèle conceptuel des connaissances, objet de ce chapitre, constitue à la fois un cadre pour comprendre et interpréter les informations provenant de l'expert et un langage pour les formaliser en vue de construire un système (Aussenac-Gilles, Krivine *et al.*, 1992). La notion de modèle conceptuel a été introduite en 1990 comme support pour décrire les connaissances des systèmes à base de connaissances (SBC) indépendamment de leurs implémentations (Newell, 1982). Il sert de support pour guider l'acquisition des connaissances. Cette même proposition de Newell a également justifié l'organisation des modèles conceptuels en plusieurs parties. Il distingue les connaissances du domaine des connaissances de raisonnement à savoir des actions et des buts, modélisés via des méthodes et des tâches.

3.1.4 Objectif

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. Ainsi, la préservation du sens de l'information échangée est alors un problème important. C'est ce que l'on appelle l'*interopérabilité sémantique*. D'après (Charlet, 2002) l'interopérabilité sémantique : « consiste à donner un sens aux informations échangées et s'assure que ce sens est commun dans tous les systèmes entre lesquels des échanges doivent être mis en oeuvre ».

L'objectif de ce chapitre est de proposer un modèle permettant de représenter les connaissances contenues dans les informations extraites dans le chapitre précédent traitant des interactions entre les facteurs de risque cardiovasculaires. La représentation des connaissances est un moyen indispensable pour dépasser la gestion des informations et entreprendre celle des connaissances. Pour (Paquette, 2002), la représentation des connaissances est au coeur de deux processus inverses et complémentaires :

1. d'abord l'extraction des connaissances dans des documents : vise à les rendre largement disponibles (sous forme d'informations) pour le travail et la formation d'autres personnes ;
2. ensuite l'acquisition de connaissances et de compétences nouvelles : vise à transformer des informations en connaissances par des personnes, au moyen d'activités

formelles ou informelles d'apprentissage utilisant une variété de formes et de supports.

Par ailleurs, une structuration des informations sous forme d'un modèle conceptuel vise à tirer certains avantages :

— des apprentissages signifiants : du point de vue de la théorie de l'apprentissage signifiant (*meaningful learning*) proposée par Ausubel (1968), la création de liens entre les connaissances est fondamentale dans le processus de construction des connaissances puisqu'une connaissance ne peut être comprise tant qu'elle n'est pas reliée de façon signifiante aux connaissances antérieures d'un individu. C'est d'ailleurs en s'appuyant sur cette théorie que les premiers usages de la carte conceptuelle en éducation ont été justifiés, dans les années 1980 (Novak *et al.*, 1984).

— une aide à la structuration des connaissances : cette structuration oblige à rendre explicites et plus précises des associations implicites et souvent confuses, un processus au coeur de la construction de significations (Fischer et Mandl, 2000).

— un outil d'amplification cognitive : pour (Jonassen et Marra, 2011), l'activité de construction d'une carte de connaissances ne réduit pas la charge cognitive requise pour traiter l'information et ne rend pas l'apprentissage plus facile, au contraire, elle oblige les apprenants à s'engager dans des processus de traitement profond des connaissances traitées.

— un moyen de favoriser la pensée réflexive : selon (Novak *et al.*, 1984) c'est un excellent exercice pour apprendre à « manier les idées » puisqu'elle permet de « manipuler » les concepts (séparation, réunion, déplacement, etc.).

3.2 Approche proposée

L'approche adoptée pour créer notre modèle conceptuel de connaissances est inspirée des travaux de (Aussenac-Gilles, Laublet *et al.*, 1996), dans lesquels les connaissances sont réparties selon la notion de point de vue en connaissances du domaine et connaissances stratégiques. Le niveau domaine décrit les concepts du domaine et leurs relations. Le niveau stratégique se base sur le niveau domaine et exprime comment une tâche va être effectuée (voir figure 3.3).

Dans notre cas les tâches identifiée sont les suivantes :

Construction du modèle conceptuel : pour cette tâche nous distinguons deux sous-tâches : (i) extraction des informations concertant le contexte de l'étude. Cette tâche a été traitée dans le chapitre précédent ; (ii) transformer et structurer les informations extraites en connaissances en s'appuyant sur des méthodes identifiées dans la littérature. Nous avons adopté l'approche proposée par (Aussenac-Gilles et Matta, 1994), qui consiste à construire le modèle conceptuel en trois étapes (voir figure 3.4) : (i) construction d'un premier modèle conceptuel appelé *Schéma du modèle conceptuel* (*framework of the conceptual model* dans CommonKADS (Schreiber *et al.*, 2000)), car il va subir des modifications au cours des étapes suivantes. Ce schéma comporte des indications sur la structuration des connaissances du domaine en précisant les concepts et la sémantique des relations qui les associe ; (ii) définition du modèle conceptuel complet ou raffinement du modèle conceptuel à partir du premier schéma ; (iii) validation de ce dernier modèle par des experts du domaine.

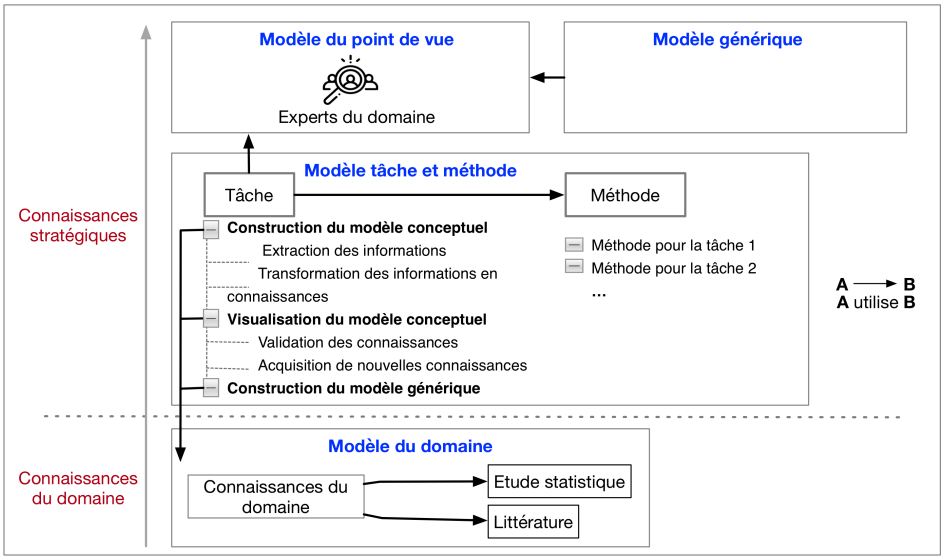


FIGURE 3.3 – Modèle conceptuel de connaissance selon la notion de point de vue

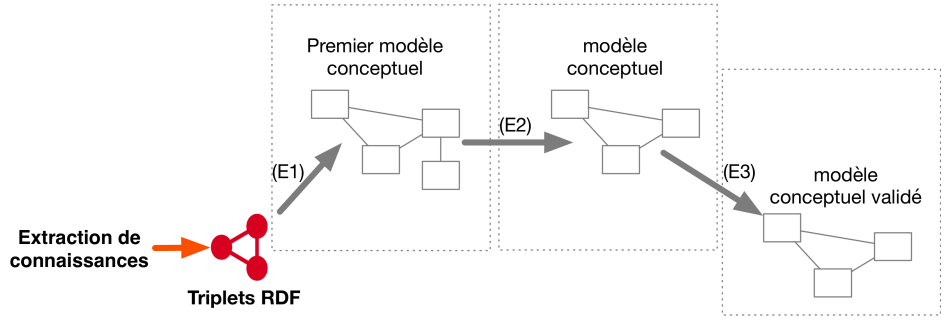


FIGURE 3.4 – Approche de construction du modèle conceptuel

Visualisation du modèle conceptuel : l'objectif de cette tâche est d'aider l'expert lors du processus de validation au niveau de la tâche précédente (plus préciciment à l'étape (E2) de la figure 3.4) et d'acquérir de nouvelles connaissances à partir des actions et des observations réalisées sur le modèle conceptuel (par exemple, valider un concept, affiner un concept, etc.).

Construction du modèle générique : l'objectif de cette dernière tâche est de construire un modèle générique de connaissances pour le domaine étudié. Souvent, divers modèles de connaissances décrivent un même domaine. Ce qui nécessite de mettre en place une démarche permettant d'articuler ces modèles pour les exploiter et guider la construction du modèle générique. La réutilisation d'éléments de connaissance déjà structurés est souvent privilégiée pour réduire le coût de la modélisation de connaissances (AUSSENAC-GILLES, CHARLET *et al.*, 2012). Ceci n'est possible que si l'on connaît bien les principes selon lesquels ces modèles ont été construits. Il est alors possible de les comparer, de les combiner, d'en sélectionner des fragments pour les recomposer au sein de nouveaux modèles.

3.2.1 E1 : construction du Premier modèle conceptuel

L'objectif de cette section l'élaboration d'un schéma du modèle conceptuel des interactions entre les facteurs de risque cardiovasculaire. Ces derniers vont être représentées par des *concepts* et des relations (ou *liens*) typées entre les concepts. Nous

avons cherché à faire émerger un premier modèle à partir des termes du vocabulaire construit dans le chapitre précédent. Ce choix d'approche découle : (i) de la volonté d'avoir une description structurée et adaptée aux fonctionnements définie précédemment ; (ii) de l'absence de modèle connu décrivant les interactions entre les FRC ; (iii) du nombre restreint de concepts décrivant les interactions entre les facteurs de risque cardiovasculaires. La liste des concepts et des relations pourra être complétée.

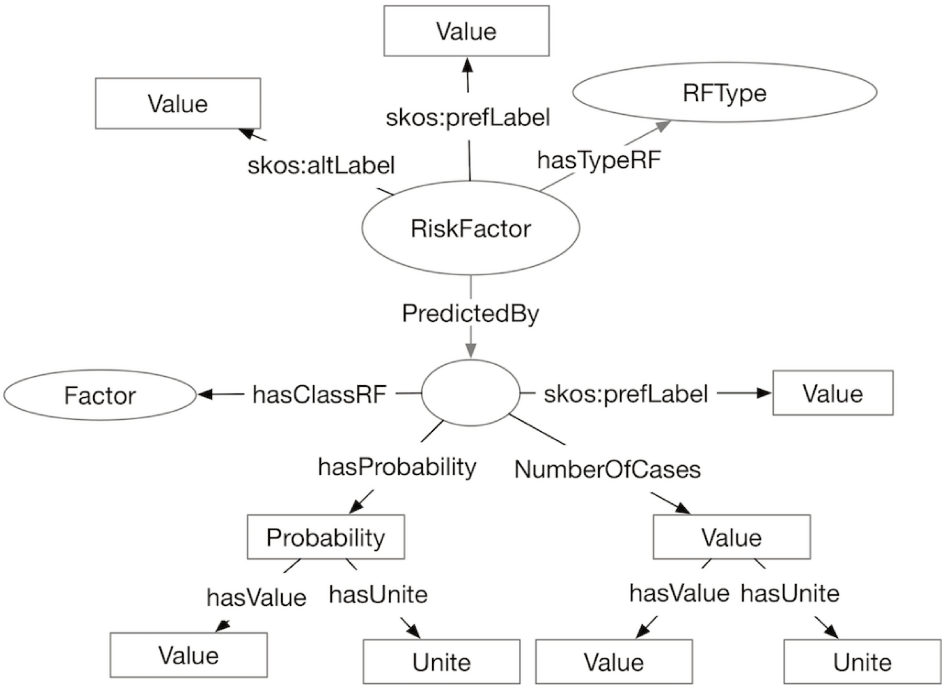


FIGURE 3.5 – Modèle conceptuel

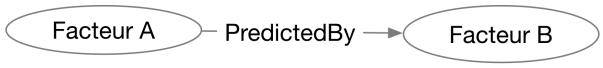
La figure 3.5) illustre le schéma du modèle conceptuel obtenu. Dans ce modèle, chaque facteur de risque est décrit par : sa catégorie, les facteurs qui le prédisent, le label préféré et le label alternatif. Par exemple, le facteur de risque « Smoking » est décrit par :

1. sa catégorie : par exemple, pour le facteur de risque « Smoking » sa catégorie correspond à « Behavioral_factors » ;
2. ses liens entretenus avec les autres facteurs : le lien « PredictedBy » définit la liste des facteurs prédisant chaque facteur. Chaque facteur prédit est modélisé sous forme d'un noeud blanc. Ce choix a été adopté dans le chapitre précédent lors de l'extraction des informations à partir des tableaux. Par exemple, le facteur de risque « Smoking » prédit par les facteurs « Obesity » et « Depression » (voir figure 3.6) ;
3. son label préféré et label alternatif. Par exemple, « Smoking » et « Fumeur » respectivement.

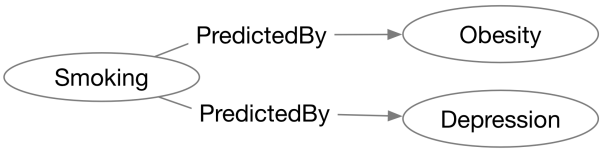
Une fois ce Schéma établi, nous avons structuré les informations extraites dans le chapitre (sous forme de triplets RDF) précédent selon ce modèle. Le résultat obtenu est un graphe RDF.

3.2.2 E2 : définition du modèle conceptuel complet

Une fois le schéma du modèle conceptuel défini, il est utilisé pour orienter le recueil de nouvelles connaissances et corriger ou instancier le schéma. L'expert peut alors



(A) Schéma du premier modèle conceptuel



(B) Description du facteur « Smoking » selon le Schéma du premier modèle conceptuel

FIGURE 3.6 – Schéma du modèle conceptuel des interactions entre les facteurs de risque cardiovasculaire

chercher à valider ou affiner les éléments du schéma pour l'enrichir ou le réviser.

3.2.2.1 Visualisation du premier modèle conceptuel

Un des principaux intérêts des modèles conceptuels est d'améliorer l'accès à la connaissance grâce aux technologies du web sémantique. Ainsi, dans l'affichage des informations, la navigation doit s'appuyer sur la sémantique. Toutefois, cette structuration sous forme de graphe est peu accessible à un non expert d'en comprendre les mécanismes. C'est pourquoi nous avons opté pour une approche de visualisation pour guider l'utilisateur pendant ce processus d'accès à la connaissance. Cette visualisation des connaissances, vise à améliorer le transfert et la création de connaissances parmi les personnes en leur donnant des moyens plus riches d'exprimer ce qu'elles savent (EPPLER et R. A. BURKHARD, 2004).

Dans ce cadre, nous avons développé l'outil MCVGraphViz (qui sera le sujet du chapitre suivant). Son principe est la visualisation dynamique de cette représentation sous forme de graphe permettant un parcours des différents noeuds guidé par les interactions existant entre ces mêmes noeuds. Dans notre cas d'étude, le parcours du graphe permet un accès aux connaissances relatives aux interactions entre FRC (voir figure 3.7).

3.2.2.2 Validation des connaissances acquises

L'objet de cette partie est d'évaluer la pertinence des connaissances extraites. Le modèle conceptuel a subi une validation préliminaire, en le confrontant à deux experts dans le domaine de l'épidémiologie. L'évaluation consiste à effectuer une analyse du modèle conceptuel, de ses entités et de ses relations. L'évaluation inclut la validation et la vérification par rapport aux connaissances du domaine.

Le processus de validation est nécessaire afin de s'assurer que les données issues de la présentation statistique sont correctement interprétées et que la transformation vers le modèle conceptuel n'a pas introduit une mauvaise interprétation de l'information. Pour valider le modèle conceptuel, nous avons eu recours à l'expertise de l'expert, mais aussi de valider le modèle à travers un environnement logiciel associé (outil MCVGraphViz).

Le processus de vérification vise à assurer que les connaissances définies dans le modèle conceptuel soient conformes à celle décrite dans la représentation statistique. Plus généralement, la vérification s'applique à chaque entité et chaque relation décrite

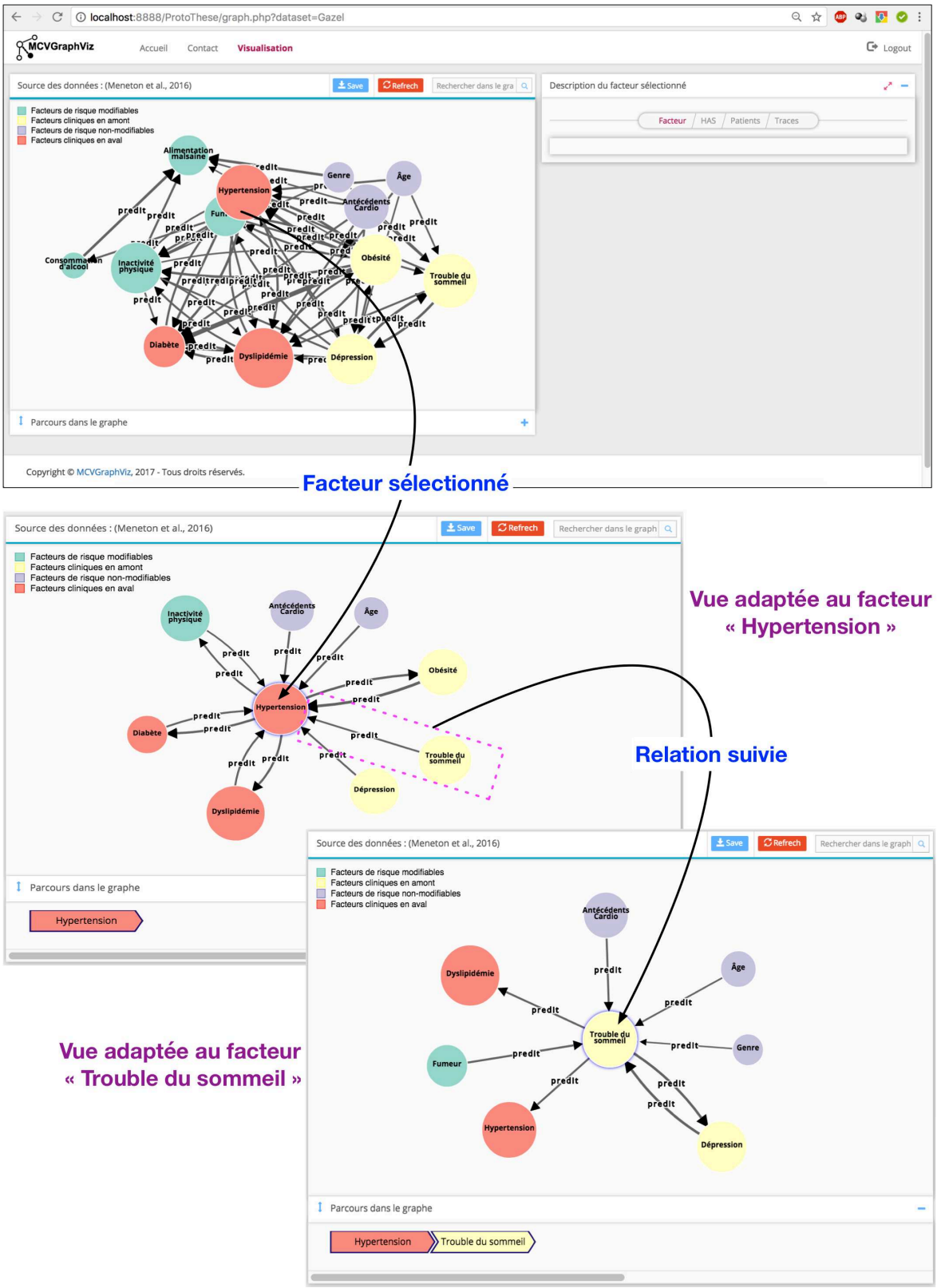


FIGURE 3.7 – Interface de l'outil MCVGraphViz avec le graphe de l'étude de (MENETON *et al.*, 2016)

dans le modèle conceptuel. Nous avons évalué avec l'expert chacun de ces éléments en nous appuyant sur l'outil MCVGraphViz. La figure 3.8 présente un exemple de vue du facteur de risque « Trouble du sommeil » avec l'ensemble des relations entretenu avec

les autres facteurs, ainsi l'expert peut discuter de la pertinence de chaque élément de ce graphe.

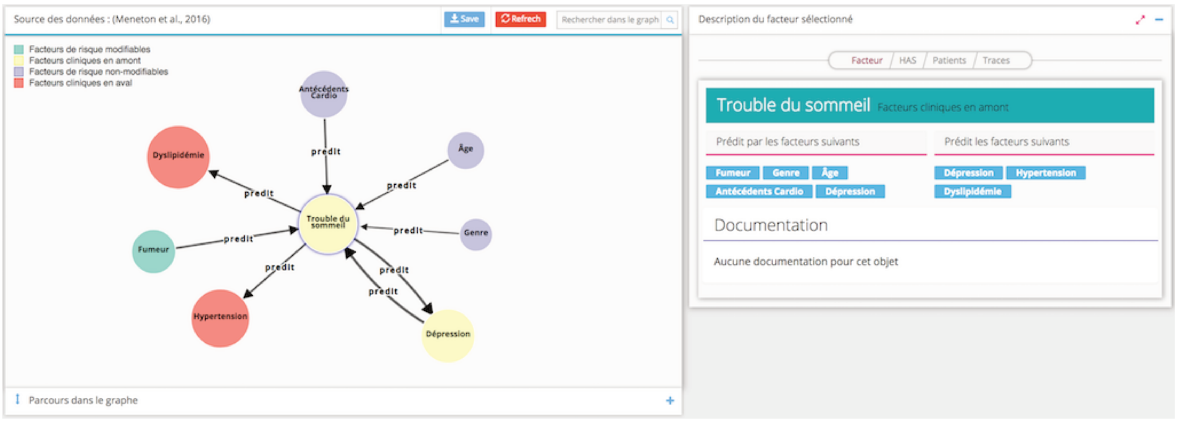


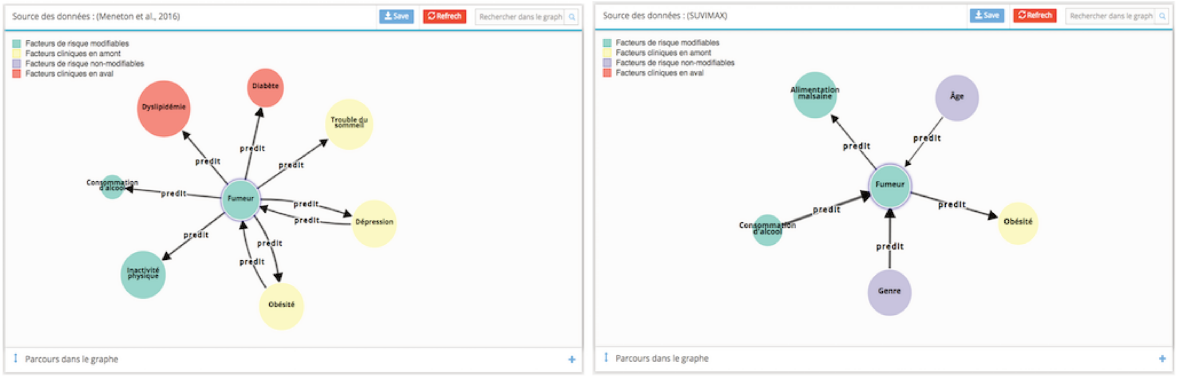
FIGURE 3.8 – Vue du facteur de risque « Trouble du sommeil » dans MCVGraphViz lors de la validation par l'expert

3.2.3 Discussion

Nous constatons ainsi que la sémantique véhiculée par chaque modèle conceptuel est plus ou moins explicite. Ces modèles conceptuels sont importants, car ils permettent de décrire les connaissances d'un domaine sous forme de concepts et de relations. Ainsi, il est possible de répondre aux questions nécessaires à la compréhension grâce au raisonnement rendu possible par ces derniers. Dans ce contexte, ces modèles vont plutôt servir de base pour caractériser les connaissances sur un domaine et construire un modèle générique en testant la validité de chaque relation obtenue dans les différents modèles en faisant appel à des connaissances du domaine soit auprès des experts, soit dans la littérature.

3.2.4 E3 : définition et validation du modèle conceptuel complet

Comme montré dans les sections précédentes, les modèles conceptuels obtenus à partir d'études différentes peuvent être différents. Pour mieux comprendre et interpréter ses variations, nous avons comparé avec les experts les deux modèles conceptuels obtenus à partir des jeux de données (D1) et (D2).



(A) Vue du facteur « Fumeur » à partir de (D1)

(B) Vue du facteur « Fumeur » à partir de (D2)

FIGURE 3.9 – Vue du facteur « Fumeur » dans MCVGraphViz à partir de (D1) et (D2)

Un des premiers objectifs de cette comparaison des modèles obtenus est de réaliser un bilan sur les connaissances traitées. Ce bilan est essentiel pour les experts. Il permet de mettre en évidence les connaissances apportées. En contrepartie, il désigne les éléments de variations entre les modèles. Ces éléments sont, soit des éléments connus de la littérature et donc leur contenu sont existants, soit des trous de connaissances scientifiques. La définition de ces trous est fondamentale pour orienter de façon pertinente de futures actions de recherche. Elle est d'autant plus fondamentale que ces travaux ont pour objectifs d'apporter les connaissances essentielles aux préoccupations de la conception.

Considérons, par exemple, l'extrait des deux modèles conceptuels ((D1) et (D2)) représenté dans la figure 3.9 et visualisé dans l'outil MCVGraphViz. Cette représentation visuelle des modèles permet, pour les experts et non experts, une lecture facile du graphe. Intuitivement, nous remarquons une variation dans le nombre de liens et de noeuds pour le même facteur de risque « Fumeur ». Si nous poursuivons la décomposition de cette figure, les experts ont identifié les points suivants : (i) certains facteurs de risque sont bien connus dans la littérature ce qui explique leurs présences dans les deux modèles conceptuels ; (ii) il y a une variation entre les deux modèles conceptuels concernant le nombre de relations et la nature des relations entre les facteurs de risque.

Les variations observées à partir de ces deux modèles sont le résultat d'une construction à partir d'études statistiques menées par différents experts ayant chacun leur propre expérience et approche. Cette situation conduit souvent à des modèles conceptuels variés.

Concernant la nature des relations qui lie deux concepts, les experts ont souligné sa nature complexe. D'après ces derniers, pour étiqueter une relation entre deux concepts il est essentiel de prendre en compte un facteur essentiel qui est l'« individu ». Plus exactement la présence ou non du facteur de risque chez ce dernier. Cette contrainte donne différentes possibilités pour l'étiquette de la relation entre deux concepts. Prenons par exemple, les deux facteurs « Fumeur » et « diabète » qui sont en relation dans la figure 3.9-(A), les étiquettes possibles pour cette relation peuvent être résumées dans la figure 3.10. Si par exemple, l'individu fume (fréquence élevée) il est 40% plus susceptible de développer le diabète par rapport à une personne non-fumeur et le fait de fumer pour une personne diabétique augmente le risque de diabète mellitus de type 2.

Dans le même ordre d'idées, les experts n'ont pas validé certaines relations par manque de littérature ou de connaissance suffisante les concernant. Par conséquent, nous sommes face à des trous de connaissances qu'il faut combler.

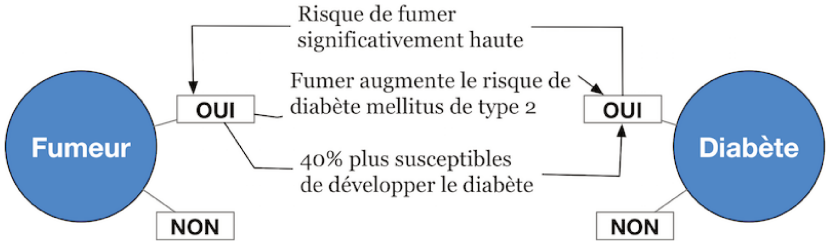


FIGURE 3.10 – Nature des liens entre « Fumeur » et « diabète » suivant la présence ou non chez un individu

La construction de ce modèle a permis de représenter les facteurs de risque et leurs interactions. Ce modèle conceptuel pourrait donc servir de cadre à une évaluation

du risque mettant en scène des utilisateurs différents. Chaque utilisateur présente des facteurs de risque dont la prise en compte des interactions est primordiale pour une meilleure évaluation du risque et pour une prise de décision efficace concernant les actions à mettre en place pour réduire ce risque.

Toutefois, ce modèle conceptuel présente également des points faibles. L'une des limites est relative à l'instanciation du modèle générique. La diversité des liens et leur nature complexe amènent en effet à s'interroger sur la robustesse de ce modèle conceptuel. Il conviendra donc dans un travail ultérieur de tester le modèle conceptuel à partir de l'instanciation de plusieurs modèles conceptuels à partir de différentes études. Cette instanciation permettra de plus de finaliser l'ontologie du domaine des interactions entre les facteurs de risques cardiovasculaires.

3.3 Conclusion

Nous avons présenté le modèle conceptuel, construit pour permettre aux acteurs de l'épidémiologie de partager leurs connaissances sur les facteurs de risque cardiovasculaire. Une autre spécificité de ce modèle conceptuel est sa finalité, celle de la conception d'un modèle de connaissances générique sur les facteurs de risque cardiovasculaire qui favorise le partage de la dimension implicite du savoir. Or, la façon dont le modèle générique décompose la réalité dépend intrinsèquement de la tâche à laquelle elle est dédiée : faciliter l'explicitation et la description de la nature des liens entre les facteurs de risque. Mais la conception d'une terminologie adaptée à l'utilisateur ainsi qu'une réflexion plus poussée sur la nature des relations à réaliser sur les graphes doivent encore être menées pour formaliser le modèle générique construit sous la forme d'une ontologie du domaine. Dans le chapitre suivant, nous présentons l'exploitation de ce modèle conceptuel dans un outil de visualisation MCVGraphViz destiné à l'évaluation du risque en prenant en compte les interactions entre les facteurs de risque.

Chapitre 4

MCVGraphViz, un outil de visualisation : application au graphe d'interaction de facteurs de risque cardiovasculaire

Sommaire

4.1 Définition et objectifs **78**
4.1.1 Cognition et perception visuelle 78
4.2 Visualisation d'information *vs* connaissances **81**
4.2.1 Définition de la visualisation 81
4.2.2 Visualisation d'information 82
4.2.3 Visualisation de connaissances 88
4.2.4 Quelle approche choisir pour le graphe de connaissances ? 94
4.3 Conception de MCVGraphViz **95**
4.3.1 Scénario d'usage 95
4.3.2 Encodage graphique 95
4.3.3 Aspect dynamique 96
4.3.4 Implémentation 97
4.4 Conclusion **101**

L'objet de ce chapitre est de présenter l'outil MCVGraphViz conçu pour visualiser le graphe des interactions entre facteurs de risque. Il s'appuie sur le modèle conceptuel construit pour fournir à l'utilisateur une interface de visualisation permettant d'accéder de façon progressive et dynamique au contenu du graphe. De plus, le système collecte les actions effectuées au cours de la navigation permettant ainsi de constituer une trace et de la stocker sous la forme d'un chemin (nœud de départ, nœud de destination).

Textes de références :

R. Azzi, S. Despres et J. Nobécourt. (2018). Visualisation dynamique de connaissances : Application aux interactions entre facteurs de risque des maladies cardiovasculaires.18eme Conférence Intérnationnale sur l'Extraction et la Gestion des Connaissances (EGC) 2018, 2p présentation poster, Jan 2018, Paris, France.

R. Azzi, S. Despres et J. Nobécourt. (2018). Visualisation dynamique de connaissances : Application aux interactions entre facteurs de risque des maladies cardiovasculaires. Atelier VIF/EGC 2018, présentation 15min sélection *via* résumé, Jan 2018, Paris, France.

4.1 Définition et objectifs

4.1.1 Cognition et perception visuelle

Nous avons choisi de commencer par un bref aperçu de l'histoire de la cognition et de la perception visuelle. Ce n'est en aucun cas un examen approfondi, mais un résumé visant à introduire ces deux concepts.

Depuis longtemps, la transmission des idées passe principalement par la sémantique exprimée par les langages textuels. Ainsi, la composition des phrases par des mots positionnés dans un certain ordre guide l'œil dans la lecture. Cependant, même si cela nous permet d'exprimer des idées complexes et subtiles, certaines représentations vont au-delà du mode linéaire et ajoutent une distribution spatiale à l'information, éventuellement sur un support interactif, ou *via* une approche combinant une vue globale et une vue détaillée (Saint-Cyr et Parade, 2018). Ces représentations ont conduit à l'émergence de la *sémiotique*[1].

En 1912, des psychologues allemands fondent la théorie de la *Gestalt* qui définit les principes de la perception. Le principe de base de cette théorie : (i) stipule que « Le tout est plus que la somme de ses parties ». Par exemple, une image est composée de points colorés (des pixels sur un écran, des taches lumineuses sur la rétine, etc.), mais contient plus que ces points : la position de chaque point *par rapport à l'ensemble des autres points* doit aussi être prise en compte. Ce sont ces positions relatives qui constituent la forme. En conséquence, une forme *Gestalt* ne peut pas être décomposée en sensations indépendantes ; (ii) les formes *Gestalt* existent en dehors de l'esprit qui la perçoit, et sont donc au moins partiellement indépendantes de l'observateur. Selon la « théorie de la Gestalt », notre cerveau va essayer de donner un sens à des formes qui n'en ont pas initialement. Par exemple dans la figure 4.1, voyons-nous un carré et quatre cercles, ou bien quatre cercles dont il manque un segment ? Ce n'est que parce que ces éléments sont parfaitement organisés et coordonnés entre eux qu'il peut en ressortir quelque chose de plus grand. Très souvent, plusieurs formes peuvent être perçues dans une image, cependant l'une d'entre elles est la bonne forme, celle qui sera perçue par l'œil. La bonne forme est indépendante de la personne qui regarde l'image. Plusieurs lois permettent de déterminer quelle est la bonne forme :

— **Loi de bonne forme :** la forme reconnue parmi un ensemble d'éléments correspond à la plus simple, tels que des groupes de points aléatoires. Par exemple, c'est ce qui arrive avec les constellations lorsque vous regardez les étoiles dans le ciel ;

— **Loi de proximité :** les éléments les plus proches vont être perçus comme appartenant à un même groupe ;

— **Loi de similarité :** les éléments ayant le plus de similarités vont induire un sens identique ;

— **Loi de symétrie :** les formes ayant un axe de symétrie sont mieux perçues que les autres ;

— **Loi de clôture :** les vides entre différentes parties d'un ensemble sont spontanément comblés ;

— **Loi de répétition :** lorsqu'une loi se répète plusieurs fois, son effet augmente.

1. La sémiotique est l'étude des signes et des systèmes de signes ; contrairement à la sémantique, elle ne se limite pas aux langages textuels et essaie d'étendre les projets de la linguistique à d'autres systèmes de signes comme les langages graphiques (Vaillant, 1999)

Ces lois permettent de guider les concepteurs d'interfaces pour prendre en compte la psychologie visuelle des utilisateurs.

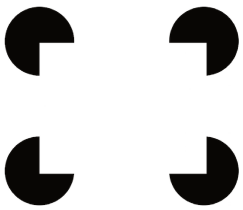


FIGURE 4.1 – Exemples d'application de la loi de *Gestalt*.

Neurath (NEURATH, 1936) est considéré comme le pionnier des systèmes de signes (ou de la conception de l'information) avec sa contribution à la conception graphique et sa proposition du système de langage visuel ISOTYPE (International System of Typographic Picture Education). Il a collaboré au projet de planification urbaine et il a appliqué ISOTYPE comme étant le langage optimal pour visualiser la ville moderne. La sémiotique des pictogrammes était théorisée afin de définir une communication visuelle par des symboles graphiques compris par tous.

Vers les années 60, Paivio (PAIVIO, 1969) propose la théorie du double codage. Il estime que le code imagé est d'autant plus utilisé que le sujet doit traiter des éléments concrets. Cette théorie est d'autant plus appuyée par les travaux de (DUCHARME et FRAISSE, 1965). Dans leurs travaux, ils ont comparé trois scénarios : (i) la présentation de dessins ; (ii) la présentation de dessins plus dénomination ; (iii) la présentation de mots. La mémoire des sujets dans les scénarios (i) et (ii) est équivalente, mais supérieure au scénario (iii). Les dessins suscitent donc les mots et sont codés de façon imagée et verbale. En tenant compte de ces résultats, les images ont la plus forte probabilité d'être codées doublement, suivie des mots concrets alors que les mots abstraits seraient codés que de façon verbale.

Plusieurs auteurs ont proposé des classifications des encodages graphiques dans le but de déterminer les encodages les plus efficaces en fonction des variables. Les travaux de (BERTIN, 1999), dont la *sémiologie graphique* est considérée comme la première théorie structurale des graphiques. Elle définit les caractéristiques et les composantes d'un graphique et leur impact cognitif. Dans ses travaux en (BERTIN, 1999), les variables graphiques étudiées sont : (i) les « variables rétiniennes » qui correspondent à la taille [2], la luminosité, la texture, la couleur, l'orientation, et la forme (voir figure 4.2) ; (ii) la variable position qui joue un rôle particulier en visualisation puisqu'il s'agit de l'information visuelle dominante dans une représentation.

Afin d'estimer les différentes possibilités d'encodage avec ces sept variables graphiques, (BERTIN, 1999) identifie quatre attitudes envisageables pour une personne face à des données :

— la perception associative, lorsque l'utilisateur cherche à regrouper les différentes modalités d'une variable nominale pour pouvoir les repérer toutes ensemble ;

— la perception sélective, lorsque l'utilisateur cherche à distinguer les différentes modalités d'une variable nominale ;

— la perception ordonnée, lorsque l'utilisateur cherche à percevoir l'ordre des modalités d'une variable ordinale ;

2. c'est la variation de surface qui constitue le stimuli sensible de la variation de taille (BERTIN, 1999)

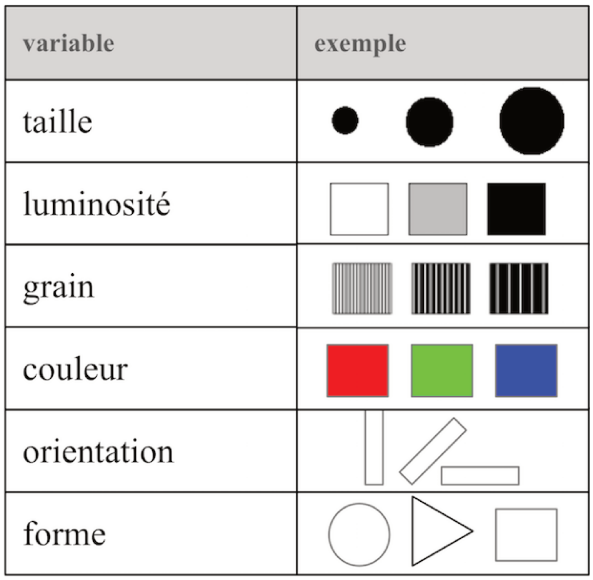


FIGURE 4.2 – Les variables retiniennes de (BERTIN, 1999).

type de variable à représenter / variable / perception mise en jeux	variable nominale		variable ordinale	variable quantitative
	associative	sélective	ordonnée	quantitative
position	+	++	+	+
taille		+	+	+
luminosité		+	+	(+)
grain	+	+	+	
couleur	+	++	(+)	
orientation	+	+		
forme	+	(+)	(+)	(+)

TABLE 4.1 – Propriétés des différentes variables selon (BERTIN, 1999). Pour les propriétés, « + » indique que la propriété est présente pour une variable, « ++ » qu'elle est très prononcée, « (+) » qu'elle peut être présente mais à certaines conditions.

— la perception quantitative, lorsque l'utilisateur cherche à percevoir les rapports entre les valeurs d'une variable quantitative.

Le tableau 4.1 synthétise ses associations en montrant l'adéquation entre variables graphiques et variables à représenter.

Différents travaux en psychologie cognitive ont mis en évidence la capacité du cerveau à analyser rapidement des composantes graphiques, et à pouvoir raisonner sur des représentations visuelles (FORTIN et ROUSSEAU, 1989). Plus récemment, (TEBOUL et DAMIER, 2017) ont proposé une présentation du fonctionnement du cerveau humain face au changement ou à la prise de décision pour surmonter ses limites et accroître sa liberté d'action dans la gestion du changement. D'après ces auteurs, les êtres humains ont la capacité de traduire immédiatement une forme graphique en un élément sémantique. De plus, la perception visuelle humaine est bien adaptée pour appréhender une situation. Les images sont, de loin, ce que le cerveau humain reconnaît et mémorise le plus facilement. Ainsi, chaque image est décomposée et analysée selon différents éléments significatifs comme la luminance, l'orientation, la couleur, etc. Notre cerveau

est en effet particulièrement efficace pour traiter rapidement une information visuelle (on admet souvent que 90% des informations qui arrivent au cerveau sont sous cette forme).

Cette présentation a montré que la plupart des théories en cognition et en perception visuelle reposent sur des modèles psychologiques pour définir un ensemble de règles (par exemple Geslate). La « règle de la bonne forme » permet de déterminer quelle sera la forme qui sera perçue dans une image donnée, parmi toutes les formes possibles que l'on pourrait y voir. Lorsque l'on construit une représentation graphique, chaque information à représenter est associée à une variable, qui doit être choisie en prenant en compte les propriétés de l'information à représenter et de la variable.

4.2 Visualisation d'information *vs* connaissances

4.2.1 Définition de la visualisation

Le domaine de la visualisation englobe de nombreuses disciplines y compris l'interaction homme-machine, la psychologie perceptuelle, les bases de données, les statistiques et l'exploration de données, la gestion des connaissances, etc. A travers la lecture d'articles relatifs à la visualisation, nous avons identifié un certain nombre de définitions :

> *« ...the conception of imagery as a **dynamic** symbolic system capable of organizing and **transforming** the **perceptual information** that we receive »* (Paivio, 1974).

> *« Visualization is a form of communication that transcends application and technological boundaries »* (Defanti *et al.*, 1989).

> *« Visualization refers to the **cognitive** functions in **visual perception**. In visualization, **pictures** combine aspects of naturalistic representation with more **formal shapes** to enhance cognitive **understanding** »* (Arnheim, 1991).

> *« Visualization is an act in which an individual establishes a strong connection between an internal construct and something to which access is gained through the senses. Such a connection can be made in either of two directions. An act of visualization may consists of any **mental construction** of objects or processes that an individual **associates with objects** or events perceived by her or him as external. Alternatively, an act of visualization may consist of the **construction**, on some external medium such as paper, chalkboard or **computer screen**, of objects or events that the individual identifies with object(s) or process(es) in her or his mind »* (Zazkis *et al.*, 1996).

> *« Visualization for the purposes of this paper refers to any **representation** of a scientific phenomena in two dimensions, three dimensions, or with an animation. Visualizations test ideas and reveal underspecified aspects of the scientific phenomena ...display new insights and **help** investigators **compare** one conjecture with another ... illustrate an idea that words cannot describe »* (Linn, 2003).

> *« Visualization is concerned with External **Representation**, the systematic and focused public **display** of information in the form of **pictures**, diagrams, tables, and the like (Tufte, 1983). It is also concerned with Internal Representation, the mental production, storage and use of an image that often (but not always...) is the result of external representation »* (Gilbert *et al.*, 2008).

« visualization as the communication of information using graphical representations. Pictures have been used as a mechanism for communication since before the formalization of written language » (Ward *et al.*, 2015).

A la lecture des précédentes définitions, nous pouvons identifier trois distinctions à prendre en compte lors d'un processus de visualisation : (i) objet de visualisation, qui peut être des images, des représentations schématiques, des animations, etc. ; (ii) visualisation introspective, qui consiste à étudier l'objet de la visualisation. Elle correspond à une construction imaginative d'une expérience visuelle possible ; (iii) visualisation interprétative, consiste à donner un sens à un objet de visualisation ou à une visualisation introspective en interprétant l'information des objets ou des introspections et en plaçant l'interprétation dans le réseau existant de croyances, d'expériences et de compréhension.

4.2.2 Visualisation d'information

4.2.2.1 Définition

La définition la plus courante de la visualisation d'informations dans les systèmes informatiques a été apportée par (Card *et al.*, 1999). Ils définissent la visualisation des informations (*« The use of computer-supported, interactive, visual representations of abstract data to amplify cognition »*) comme l'utilisation de représentations visuelles interactives et assistées par ordinateur de données pour amplifier la cognition. L'objectif de cette visualisation est :

— de rendre l'information plus accessible en mettant en place des stratégies d'exploration par niveaux sur de grands volumes de données ;

— d'exploiter les apports de la visualisation (segmentation, reconnaissance de motifs ou de formes, etc.) pour permettre l'analyse visuelle des données.

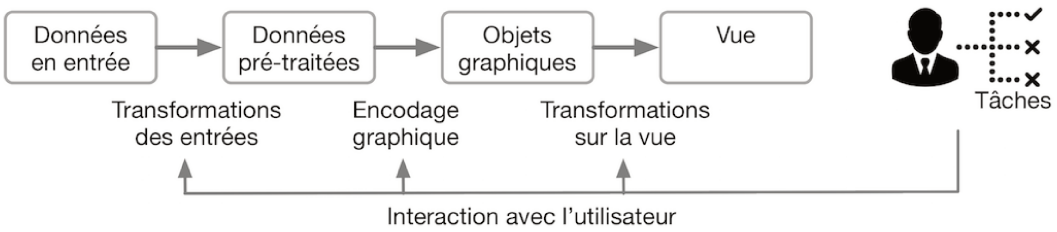


Figure 4.3 – Modèle de (Card *et al.*, 1999) pour la visualisation d'information.

La figure 4.3 illustre le modèle de référence pour la visualisation d'information proposé par (Card *et al.*, 1999). Il fournit un processus de transformation des données à partir des données brutes aux vues impliquant l'homme dans le traitement de l'interaction. Le modèle de référence est un excellent moyen de comprendre, de définir et de distinguer la visualisation des informations. Dans ce modèle, nous pouvons distinguer les processus suivants :

1. **la transformation des données**, qui correspond au traitement effectué sur les données avant leurs visualisations. Le plus souvent, les traitements concerne : (i) filtrage dynamique, qui consiste à sélectionner les éléments à visualiser par des requêtes dynamiques (l'affichage est mis à jour instantanément). Le but est de répondre à la limite liée à la visualisation de l'ensemble de données sur une

seule visualisation qui surcharge souvent les capacités de perception humaine et d'acquisition d'informations ; (ii) le détail à la demande, qui consiste à choisir un élément dans la représentation et faire apparaître des informations supplémentaires le concernant (souvent sous forme de fenêtre *pop-up* ou infobulle) ; (iii) *brushing*, qui consiste à inclure ou exclure de la visualisation de tout un sous-ensemble d'individus sélectionnés par l'utilisateur à l'aide du pointeur (Hong, 2003).

2. **l'encodage graphique**, qui consiste à intégrer les données dans une représentation graphique en associant à chaque variable de donnée une variable graphique (couleur, position, forme, etc.).
3. **la transformation sur la vue**, qui consiste à manipuler la vue proposée à l'utilisateur. Les systèmes de visualisation d'informations actuels n'offrent pas seulement une image statique. Selon (M. A. Hearst, 1999), la plupart des techniques de visualisation existantes sont fondées sur des techniques consistant à : (i) contrôler le point de vue dans l'interface par rotation ou zoom pour affiner la zone d'intérêt en déplaçant la vue sur la zone d'intérêt ; (ii) accéder à des points de vue multiples, pour permettre d'avoir une vue globale et une vue détaillée pour un même élément (*linking and brushing*) (Hiemstra *et al.*, 2009). La mise en évidence s'effectue selon différentes manières, par exemple en modifiant la couleur ou la taille des objets ; (iii) intégrer le contexte dans la vue globale et révéler les détails autour du focus (zoom). Cette technique fait face au problème lié au zoom lors du processus d'accès au détail sur la partie zoomée. Plus le facteur de zoom est élevé, plus il est possible d'afficher des détails sur des éléments particuliers mais la structure globale et le contexte des informations se perdent avec le facteur de zoom croissant (Burkhardt *et al.*, 2014). Dans certains domaines, on peut combiner plusieurs types de transformation pour répondre à un besoin. Par exemple, afficher un niveau de détail de plus en plus précis avec un zoom continuellement tant que le niveau de détail le plus faible n'est pas atteint ; (iiii) intégrer des animations, pour améliorer l'interaction (Robertson et Mackinlay, 1993) des utilisateurs en montrant les changements visuels facilement identifiables dans l'interface.

4.2.2.2 Méthode de visualisation d'information

Différentes techniques de visualisation d'information ont été proposées au cours des dernières années. Ils sont généralement classés selon la structure des données ou des informations à visualiser (texte, graphique, arbre, réseau, etc.) (Ward *et al.*, 2015). (Cao et Cui, 2016) ont proposé une revue systématique de nombreuses techniques avancées pour l'analyse de grands ensembles de documents, allant de l'élémentaire au profond, couvrant tous les aspects de la visualisation de documents texte. (Dzemyda *et al.*, 2013) proposent un état de l'art sur les techniques de visualisation de données multidimensionnelles, notamment ses applications en sciences sociales (économie, éducation, politique, psychologie), en environnement et en médecine (ophtalmologie, médecine sportive, pharmacologie, médecine du sommeil). (Long *et al.*, 2017) proposent une discussion sur les techniques de visualisation des structures en arbres. (Frati et Ma, 2018), (Kerren *et al.*, 2014) et (Tarawaneh *et al.*, 2012) proposent quelques techniques pour la visualisation des graphes et des réseaux. Enfin, des auteurs comme (Ward *et al.*, 2015), (C. h. Chen *et al.*, 2008) proposent une liste quasi exhaustive des techniques de visualisation d'information existantes.

De nombreuses techniques permettent de représenter des informations relationnelles. Par ailleurs, les représentations sous forme d'arbre, de graphe et de réseau font partie des techniques les plus adaptées pour les informations relationnelles. Toutefois, cette structuration implique la gestion de deux paramètres : (i) la nature des relations qui peuvent être simples ou complexes (par exemple, unidirectionnelles ou bidirectionnelles) ; (ii) les relations peuvent fournir des informations plus nombreuses et plus riches que celles contenues dans les données d'entrées.

Nous proposons dans la suite de cette section quelques techniques permettant de représenter les informations relationnelles.

Les arbres ou les hiérarchies : Les arbres sont parmi les structures les plus usuelles pour structurer les informations relationnelles. Nous les trouvons dans de nombreux contextes de la vie quotidienne, tels que les arbres généalogiques, les arborescences de répertoires, etc. Les arbres sont représentés par la composition de trois éléments : (i) un élément supérieur qui correspond au *nœud racine* ; (ii) des éléments intermédiaires qui sont les *nœuds internes* ; (iii) des éléments bas correspondant aux feuilles. La figure 4.4 présente cette structuration. Chaque nœud (à l'exception de la racine) a un parent et admet zéro ou plusieurs nœuds fils.

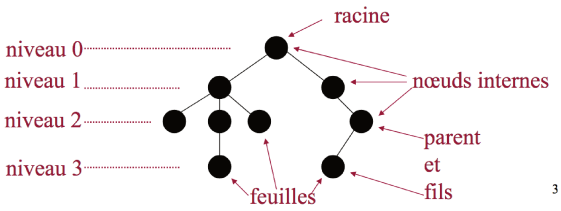


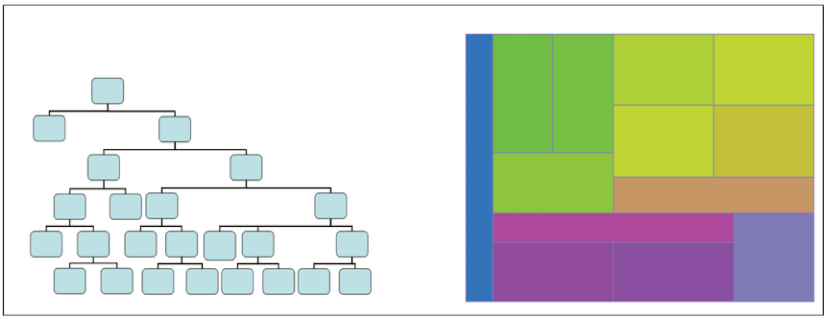
FIGURE 4.4 – Structure d'arbre

La figure 4.5 montre qu'il existe 284 techniques de visualisation pour les arbres. Selon (WARD *et al.*, 2015), ces techniques peuvent être divisés en deux catégories suivant les algorithmes utilisés :

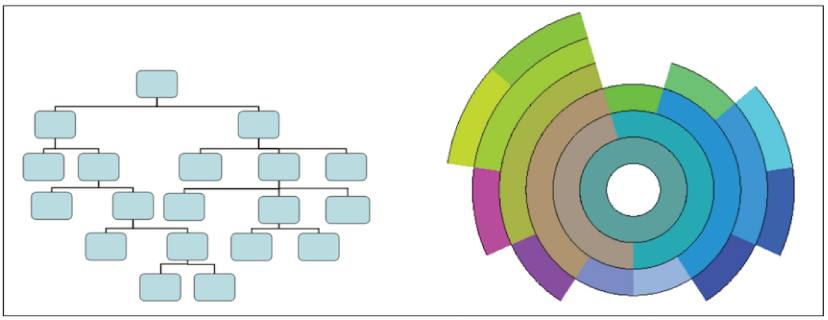
FIGURE 4.5 – Capture d'écran représentant nombre des techniques de visualisation d'arbres à cette date (source : `https://keshif.me/demo/treevis.html`)

1. pour le remplissage d'espace disponible : les deux approches les plus courantes pour générer des hiérarchies de remplissage d'espace sont les mises en

page rectangulaires ou *Treemap* (Shneiderman, 1992) et radiales ou *sunburst* (voir figure 4.6). Dans *Treemap* les nœuds sont représentés sous forme de rectangles, chaque rectangle est subdivisé en nombre de sous-rectangles égal à son nombre d'enfants. Le processus de subdivision est effectué récursivement (Schulz, 2011). Cette technique est populaire, car elle utilise efficacement l'espace de l'écran et affiche la taille des feuilles dans un arbre. Cependant, cette technique ne permet pas de visualiser la structure hiérarchique de l'arbre. L'approche *sunburst* consiste à construire une structure radiale ou circulaire à partir des relations de la hiérarchie tout en exploitant l'espace de l'écran (Stasko et Zhang, 2000). Cependant, cette structuration rend difficile la distinction des relations « parent-enfant » et le contrôle de la superficie des nœuds qui dépend du nombre d'enfants.



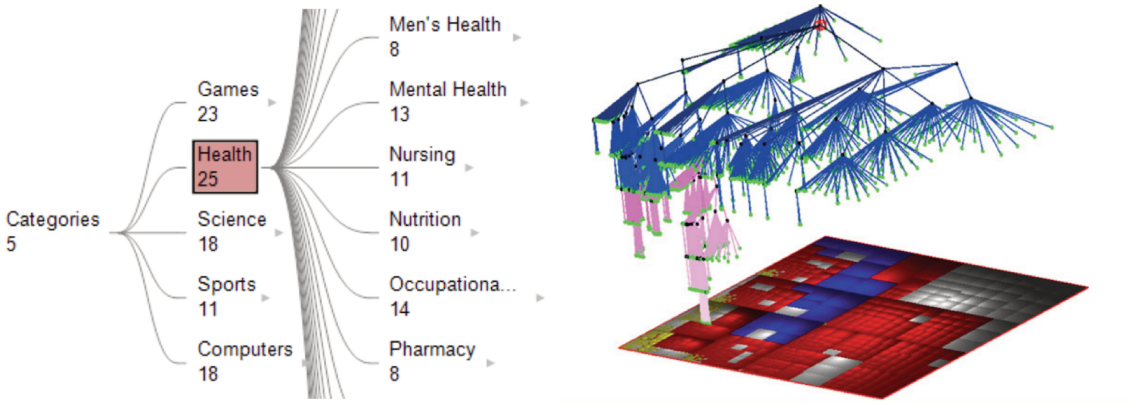
(A) Hiérarchie et sa correspondance en visualisation *treemap*



(B) Hiérarchie et sa correspondance en visualisation *sunburst*

Figure 4.6 – Exemple de remplissage d'espace disponible (Ward *et al.*, 2015).

2. pour exploiter la structure arborescente des données : la représentation la plus commune utilisée pour visualiser des relations arborescentes ou hiérarchiques, est un diagramme de liaison de nœuds dans lesquels les relations parent-enfant sont représentées comme des arêtes. Lors de la conception d'un algorithme pour dessiner un diagramme de nœud-lien, il faut respecter trois catégories de directives souvent contradictoires (Ward *et al.*, 2015), (Tarawaneh *et al.*, 2012) : (i) les conventions de dessin, par exemple la limitation des bords pour qu'ils soient une seule ligne droite ; (ii) les contraintes, par exemple exiger qu'un nœud particulier soit au centre de l'affichage ; (iii) l'esthétique, par exemple minimiser les croisements de lignes et répartir les liens et les nœuds d'une manière homogène. La figure 4.7 montre quelques exemples de méthodes de représentation de ces structures. Les chercheurs ont exploré et publié presque tous les moyens d'organiser les nœuds d'un arbre en 2D et en 3D pour les encoder sous différentes formes et les plier, les déplier ou les manipuler de manière interactive.



(a) Visualisation d'un arbre exploitant la structure en 2D (Song *et al.*, 2010)

(b) Visualisation d'un arboresence de fichier en 3D (Linsen et Behrendt, 2011)

Figure 4.7 – Exemple de représentation exploitant la structure arborescente des donneés

Les graphes et les réseaux : la visualisation de graphe est généralement fondée sur les représentations de type **Diagrammes nœuds-Liens**, ou **Diagrammes Orientés Matrices**.

1. **Diagrammes nœuds-Liens.** Un graphe est un ensemble de points que l'on appelle des nœuds reliés par des arcs. L'ensemble des arêtes (*edges* en anglais) entre les nœuds (*nodes* en anglais) forme une figure similaire à un réseau (Aggarwal et H. Wang, 2010). La plupart des techniques de visualisation développées sont fondées sur des algorithmes dont l'objectif est d'améliorer le rendu de visualisation. L'objectif est de rendre les informations plus faciles à lire, à comprendre et à manipuler. Différents travaux se sont ainsi attachés à l'optimisation de ces différents critères et cela avec différentes approches. L'un de ces critères est celui de la réduction du nombre de croisements d'arêtes. Les algorithmes par modèles de contraintes tentent de trouver un positionnement des nœuds permettant de minimiser une ou plusieurs fonctions. Le positionnement par modèle de forces (Kao, 2016),(Alhajj et Rokne, 2018) en est une version où la fonction à minimiser. Elle représente l'énergie d'un modèle physique (voir figure 4.8).

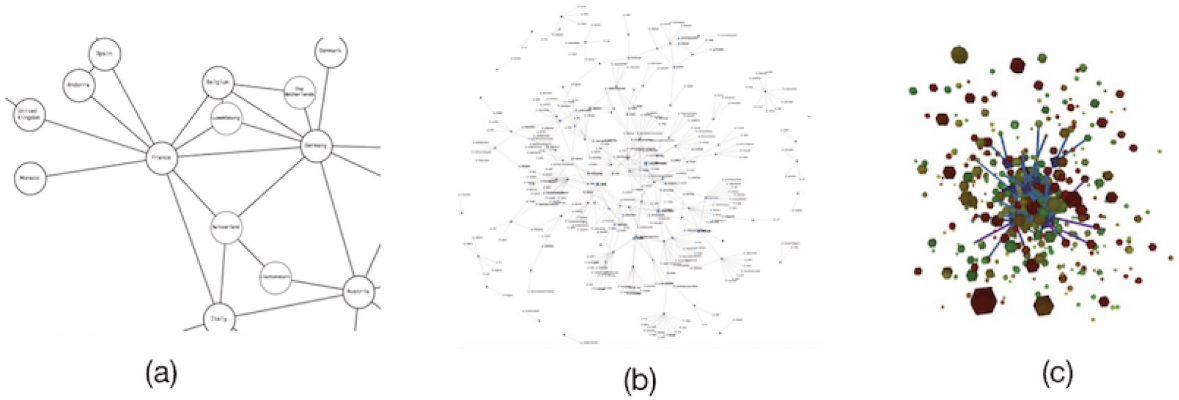


Figure 4.8 – Exemple de graphe dirigé par la force : (a) issue de (Ward *et al.*, 2015) ; (b) issue de (Rafelsberger, 2013) ; (c) issue de (Paananen et Wong, 2009).

Les méthodes et les algorithmes disponibles pour créer des graphiques dirigés

par la force exigent beaucoup de ressources par rapport aux autres méthodes de visualisation. Cela s'applique encore plus à la visualisation des systèmes dynamiques et à l'évolution temporelle des graphes. Ainsi, une étape du processus de raffinement de la position calcule la somme de toutes les forces sur chaque nœud (composantes x, y et z, selon le cas) et déplace sa position proportionnellement à cette force. De toute évidence, une fois les points déplacés, toutes les forces doivent être recalculées et un autre changement de position doit être effectué.

2. **Diagrammes Orientés Matrices.** Une bonne alternative pour la représentation visuelle d'un graphe consiste à visualiser une matrice de contiguïté (voir figure 4.9), qui est une grille N par N (où N est le nombre de nœuds), où la position *(i, j)* représente l'existence (ou non) d'un lien entre nœuds i et j. Cela peut être une matrice binaire où la valeur peut représenter la force ou le poids du lien entre les deux nœuds. Cette méthode surmonte l'un des plus gros problèmes des diagrammes nœud-lien, à savoir celui du croisement des arêtes, bien qu'il ne s'adapte pas bien aux graphes comportant un grand nombre (milliers) de nœuds. (Bertin, 1999) a été l'un des premiers à étudier le pouvoir de cette représentation. Pour cela, il a utilisé différentes stratégies de réorganisation des lignes et des colonnes afin de révéler des structures dans le graphique.

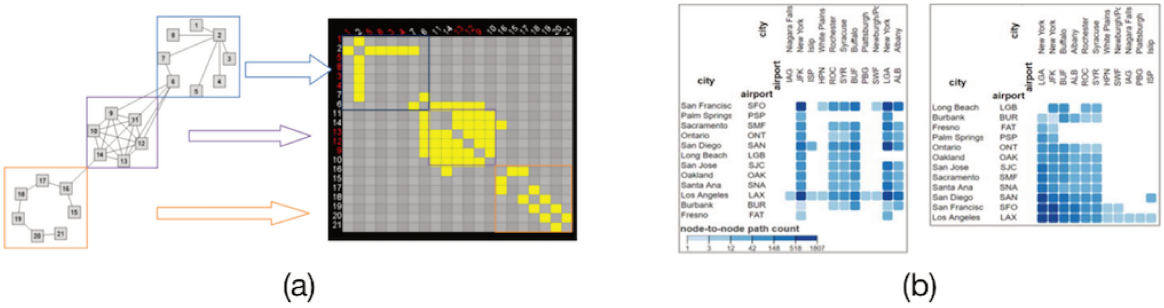


Figure 4.9 – Exemple de graphe représenté par une matrice : (a) d'après (H. Li *et al.*, 2009) et (b) d'après (Kerzner *et al.*, 2017).

Différentes améliorations ont été apportées à cette technique, notamment pour améliorer la réorganisation des lignes et les colonnes de la matrice afin d'exposer le plus de structures possibles (Perin *et al.*, 2014). Certains sont principalement axés sur l'utilisateur (Kerzner *et al.*, 2017), ce qui prend en charge le classement en fonction des valeurs de l'une des lignes ou des colonnes en tant que point de départ. D'autres sont purement automatiques et reposent sur des mesures pour évaluer un ordre particulier et une stratégie pour générer des ordres à tester (McGraw, 2015).

Les classifications habituelles reposent sur la structure des informations visualisées. Lorsque l'on considère la visualisation d'éléments et des relations entre éléments, l'un des modèles les plus utilisés est le modèle graphe qui représente les éléments par des sommets et les relations par des arêtes reliant ces sommets.

4.2.3 Visualisation de connaissances

4.2.3.1 Définition

La visualisation de connaissances est un domaine de recherche relativement nouveau. Il s'intéresse à la création et le transfert de connaissances par visualisations avec et sans l'aide d'ordinateurs. La visualisation des connaissances peine à émerger comme un domaine de recherche à part entière. Nous voyons plusieurs raisons à cela : l'absence de définition claire du domaine, les recoupements importants qui existent avec d'autres domaines comme la visualisation d'information, la représentation des connaissances ou la communication, et le fait que certaines approches visuelles n'utilisent pas d'ordinateur, ce qui rend leur appartenance à la discipline informatique moins évidente (Lamy, 2017). Par conséquent, il semble difficile de trouver une définition unificatrice à la visualisation des connaissances. Cependant, voici quelques-unes de ces définitions identifiées dans la littérature :

1) *« Knowledge visualization examines the use of visual representations to improve the transfer of knowledge between at least two persons or group of persons »* (R. Burkhard, 2004) ;
2) *« Knowledge visualization is a generic technology that investigates the use of visual representation techniques from all visualization domains that helps in knowledge-intense processes such as to create and to transfer knowledge among people »* (Z. Chen *et al.*, 2008) ;

Ensuite, un avis de huit experts sur la visualisation des connaissances a été rassemblé dans l'article (Bertschi *et al.*, 2011). Les experts avaient réfléchi à l'état actuel ou futur de ce que les auteurs considèrent comme la visualisation de connaissances. Voici quelques-unes de ces définitions :

4) la première définition est celle proposée par Tom Crawford : *« The process of Knowledge Visualization contains steps such as gathering, interpreting, developing an understanding, organizing, designing, and communicating the information. While the previous sentence implies a linear flow, it is hardly a linear process. In fact, any step can link to any other step in any number of iterations until a "final" representation is created. Of course, it never is quite final. Even if we were able to come up with the perfect visualization, our knowledge changes and with it changes the visual. In fact the visual often provides so much insight that our knowledge changes and thus the visual needs to change creating a fascinating recursive loop »* ;
5) la seconde définition est celle proposée par Andrew Vande Moere : *« Knowledge is information that has been made part of a specific context. In order for information to transform into knowledge, one must share some context, some meaning, in order to become encoded and connected to preexisting experience. In that sense, Knowledge Visualization can be considered as data visualization "in context" »* ;
6) la troisième définition est celle proposée par Sabrina Bresciani : *« Visualizing knowledge means mapping concepts graphically, by structuring text and visuals in a meaningful way. Visual representations are used to organize information and concepts in order to convey knowledge, to amplify cognition and to enhance communication »* ;
7) la quatrième définition est celle proposée par Wolfgang Kienreich : *« Knowledge Visualization is the missing link between the expert tools developed in the thriving*

field of Visual Analytics and the frantic demand of the general public for intuitive ways to cope with increasingly large and complex personal digital universes » ;

Plus récemment, (Ursyn, 2015) propose la définition suivante : *« Uses visual representations to transfer insights and create new knowledge in the process of communicating different visual formats. »*.

4.2.3.2 Méthodes de visualisation des connaissances

La visualisation des connaissances constitue un élément essentiel de la gestion des connaissances, elle vise à soutenir le processus de transfert et de création de nouvelles connaissances en utilisant des techniques de visualisation. Ces visualisations interviennent dans la vie de tous les jours et pas nécessairement sur un support informatique. Par exemple, (R. Burkhard, 2004) a observé que les architectes combinent différents types de visualisation complémentaires pour répondre à l'esprit de chaque participant [3] avec différents niveaux de détail. Concernant, la visualisation sur support informatique, de nombreuses techniques ont émergé ces dernières années pour répondre à des besoins différents. Ces techniques tiennent pour la plupart leurs origines de celles développées dans la communauté de la visualisation d'information (Herman *et al.*, 2000).

Les technologies du web sémantique offrent de nouveaux moyens d'accéder aux données et d'acquérir des connaissances. Même si l'objectif de la recherche dans ce domaine est d'avantage de fournir des données « lisibles par machine », les systèmes centrés sur l'homme bénéficient également de ces technologies. La visualisation et le Web sémantique peuvent se compléter sur un certain nombre de questions fondamentales concernant l'organisation et l'accès à des ressources d'information. L'apport de la sémantique réside dans la formalisation des connaissances du domaine. Évidemment, les représentations graphiques des données structurées jouent un rôle clé dans la recherche actuelle (Nazemi *et al.*, 2015). Dans ce contexte, nous allons présenter les approches de visualisation de connaissances.

Les cartes de connaissances. Une carte de connaissances (voir figure 4.10) est une représentation graphique d'un ensemble de connaissances d'un domaine. Contrairement au texte, les connaissances sont réparties dans un espace non linéaire sous la forme d'un réseau graphique. En effet, dans une carte de connaissances, les connaissances sont représentées de manière schématique et bidimensionnelle sous la forme d'un réseau de nœuds et d'arcs. Les nœuds représentent les idées importantes ou les connaissances, et les arcs représentent les relations (ou liens) que le concepteur de la carte établit entre les connaissances.

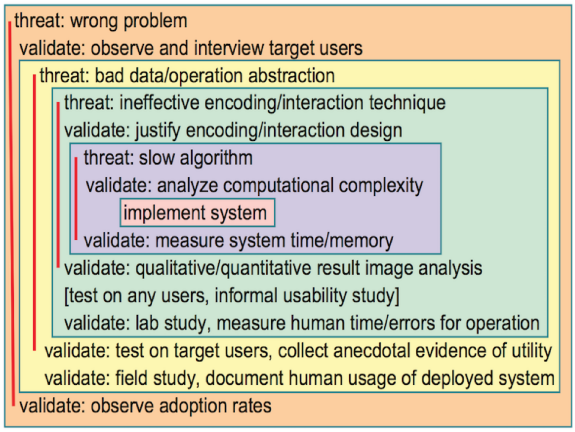
Les connaissances (ou nœuds) sont représentées au moyen d'un mot ou d'un ensemble de mots, alors que les relations (ou arcs) sont représentées au moyen de traits simples ou fléchés de manière à montrer des changements, des interrelations et aident à concevoir des stratégies (Ursyn, 2014). On peut aussi placer des étiquettes textuelles sur les traits, de manière à spécifier davantage la nature des liens établis entre les connaissances.

Il existe de nombreux outils pour la conception de carte de connaissances, nous n'avons pas pour objectif de les présenter. Cependant, le site MIND-MAPPING [4] recense plus de 450 outils.

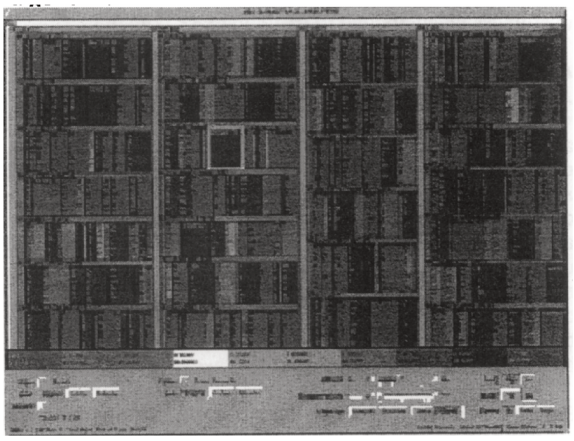
3. les ingénieurs, les travailleurs, les avocats ou les clients, qui ont des conceptions différentes et des connaissances différentes.

4. `https://www.mind-mapping.org/`

FIGURE 4.10 – Exemple de carte de connaissances (source : `https://imindmap.com/gallery/`).



(a) Exemple de visualisation en structures imbriquées (MUNZNER, 2009)



(b) Exemple de visualisation en treemap (TURO, 2003)

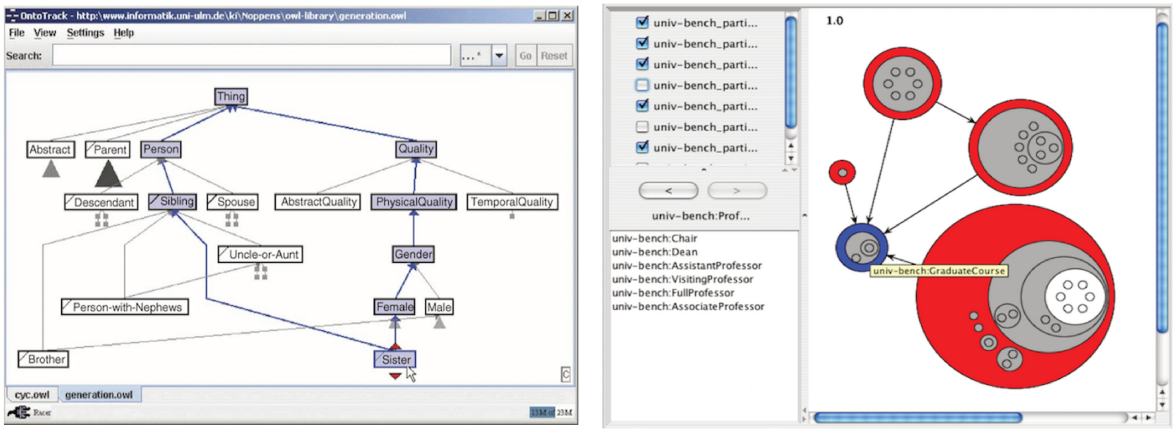
FIGURE 4.11 – Techniques de visualisation exploitant la sémantique hiérarchique

La visualisation sémantique hiérarchique.

Les structures hiérarchiques sont souvent employées pour la structuration des connaissances dans différents domaines. Même avec des structures hiérarchiques légères, une sémantique est déjà observée ce qui permet de fournir une vue compréhensible de la connaissance décrite. Ainsi, les hiérarchies offrent la possibilité de catégoriser des ressources spécifiques à un domaine dans des concepts hérités et autorisent un accès à la rubrique lié aux connaissances du domaine modélisé. Ceci permet à l'utilisateur de localiser l'information recherchée dans la structure hiérarchique et d'obtenir ainsi un point de départ pour résumer sa requête en interrogeant le nœud parent ou préciser sa recherche en choisissant un sous-nœud.

Les hiérarchies peuvent être visualisées de différentes manières, les techniques souvent employées de manière intuitive comme hiérarchique (voir figure 4.11) sont les structures : (i) imbriquées (MUNZNER, 2009) ; (ii) en treemap (TURO, 2003), (SHNEIDERMAN, 1996) ; (iii) indentées (STOREY *et al.*, 2001).

Les approches de visualisation des structures hiérarchiques utilisées dans les éditeurs d'ontologie telle que Protégé (NOY *et al.*, 2000), OntoEdit (SURE *et al.*, 2002), OntoTrack (LIEBIG et NOPPENS, 2005) sont également fondées sur des



(a) Interface de visualisation dans OntoTrack (Liebig et Noppens, 2005)

(b) Interface de visualisation dans CropCircles (T. D. Wang et Parsia, 2006)

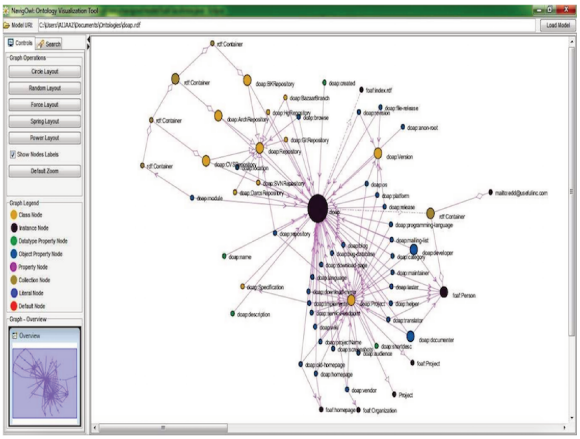
Figure 4.12 – Exemple de visualisation exploitant la sémantique hiérarchique

arbres. Considérons, par exemple, le cas de OntoTrack l'outil destiné à la visualisation et l'édition des ontologies formalisées en OWL (Antoniou et Harmelen, 2004), la visualisation est fondée sur des graphes avec une sémantique hiérarchique (voir figure 4.12(a)). Un autre exemple de visualisation imbriquée est celui de l'ontologie CropCircles (T. D. Wang et Parsia, 2006), qui représente l'arborescence de la hiérarchie de classes en tant qu'ensemble de cercles imbriqués concentriques (voir figure 4.12(b)). Cependant, une des limites de ses visualisations hiérarchiques est que la tâche principale d'acquisition d'informations par l'utilisateur en utilisant une recherche exploratoire ne soit pas activement soutenue par les systèmes introduits *supra*.

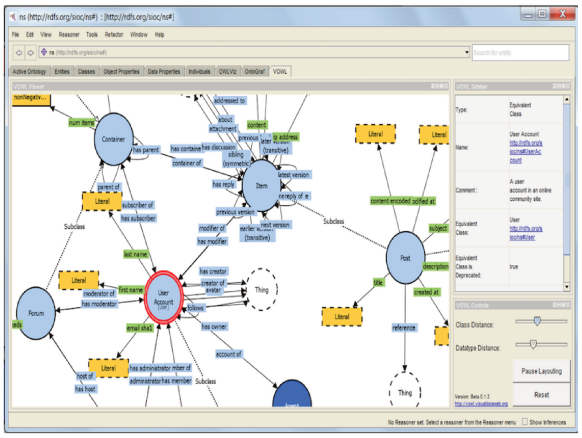
La visualisation sémantique relationnelle. Les visualisations relationnelles visent à visualiser le contexte sémantique de l'information et permettre à l'utilisateur de naviguer au sein de ce contexte. Le principe de la plupart de ces approches est d'exploiter la structure du graphe pour naviguer d'un nœud à l'autre. Dans ce contexte, divers outils ou plugins de visualisation ontologiques sont proposés. Dans le cas de l'éditeur d'ontologie Protégé divers plugins ont été proposés. OntoViz [5] plugin de Protégé, fondé sur la technologie de visualisation de graphes GraphViz d'AT&T (Peng *et al.*, 2004). OntoViz visualise la structure d'une ontologie formelle en tant que graphe arbitraire, où chaque nœud représente soit une classe, soit une instance. Il permet de visualiser uniquement certaines parties d'une ontologie en sélectionnant un ensemble de classes. L'utilisateur peut définir des couleurs pour les nœuds et les arêtes. TGVizTab (Alani, 2003) plugin Protégé fondé sur la technologie TouchGraph (Sharko *et al.*, 2008) pour visualiser graphiquement les ontologies de Protégé. L'interface utilisateur de TGVizTab consiste en un schéma d'ontologie qui illustre les hiérarchies de classes en tant que liste indentée fournie par le navigateur de classes Protégé, qui est le point d'interaction initial. Enfin, Hussain a proposé NavigOwl (voir figure 4.13(a)) (Hussain, Latif *et al.*, 2014) et Lohmann (Lohmann *et al.*, 2016) VOWL, qui vise à fournir un langage visuel complet et compréhensible pour les utilisateurs moins familiarisés avec les ontologies OWL (voir figure 4.13(b)).

(Fluit *et al.*, 2006) introduisent l'approche de visualisation Cluster Map pour

5. `https://protegewiki.stanford.edu/wiki/OntoViz`

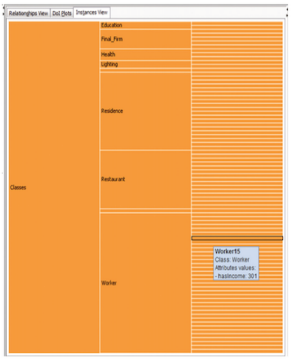


(a) Interface de visualisation dans NavigOwl (Hussain, Latif *et al.*, 2014)

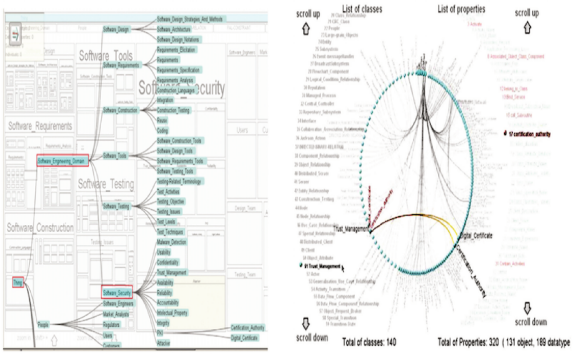


(b) Interface de visualisation dans VOWL (Lohmann *et al.*, 2016)

Figure 4.13 – Exemple de plugin de visualisation pour Protégé.



(a) Interface de visualisation dans OntoViewer (Silva *et al.*, 2012)



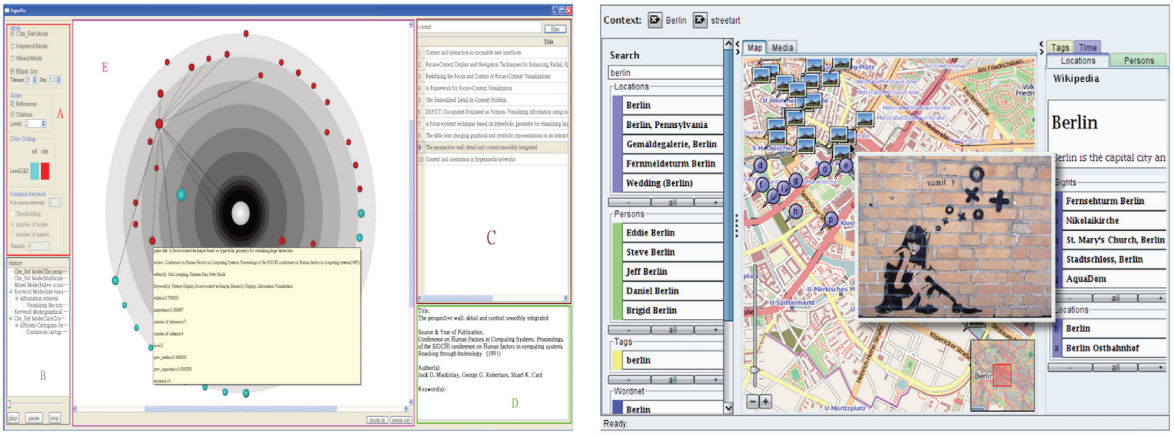
(b) Interface de visualisation dans OWL-VisMod (Garcia-Penalvo *et al.*, 2012)

Figure 4.14 – Exemple d'outil de visualisation exploitant la sémantique relationnelle

visualiser les instances d'un ensemble de classes sélectionnées en fonction de leur hiérarchie et organisées selon leurs classifications. Par conséquent, ils regroupent les instances de la même classe en un cluster groupé en forme de grappes. Chaque classe d'une carte de cluster peut avoir une couleur qui différencie la classe des autres. Le chevauchement des classes est illustré par un chevauchement des couleurs.

Une autre technique proposée pour faciliter la compréhension de la vue d'ensemble de l'ontologie (c'est-à-dire les concepts et leurs relations) est implémentée dans OntoViewer (Silva *et al.*, 2012). Cet outil facilite l'exploration visuelle de grandes ontologies en utilisant trois vues (voir figure 4.14(a)) en intégrant : (i) un arbre hyperbolique pour représenter la hiérarchie de l'ontologie ; (ii) une vue arborescente classique pour montrer les entités ontologiques ; (iii) une arborescence radiale augmentée pour afficher les relations entre les classes.

OWL-VisMod de (Garcia-Penalvo *et al.*, 2012) qui incorpore des *treemaps* pour représenter des ontologies, des arborescences hiérarchiques pour analyser la hiérarchie des concepts et des schémas radiaux pour représenter le couplage global d'une ontologie. OWL-VisMod utilise des dispositions radiales pour visualiser les relations entre les classes de l'ontologie. D'un côté du cercle se trouve une liste



(a) Interface de visualisation dans PaperVis (Panchenko *et al.*, 2013)

(b) Interface de visualisation dans SemaPlorer (Schenk *et al.*, 2009)

Figure 4.15 – Exemple d'outil de visualisation exploitant la sémantique fondée sur les entités

de classes, de l'autre une liste de propriétés. L'utilisateur a la possibilité de se concentrer sur une classe particulière (lorsqu'une classe est sélectionnée, les propriétés qui la définissent seront mises en évidence et les liens seront colorés)(voir figure 4.14(b)).

La visualisation sémantique fondée sur les entités. Contrairement aux visualisations relationnelles ou hiérarchiques utilisant la structure sous-jacente des informations sémantiques, la visualisation des connaissances se concentre sur l'ensemble des entités représentées. Elle proposent une approche exploratoire en naviguant dans une structure visuelle ou en récupérant des informations. Cette classe de visualisation de connaissances est conçue pour les tâches de recherche (Lohmann *et al.*, 2016).

De nombreuses approches ont été proposées. Par exemple, (Panchenko *et al.*, 2013) proposent PaperVis, pour aider les utilisateurs à saisir rapidement les structures de référence de citation complexe dans un groupe d'articles. PaperVis permet la recherche et l'exploration dans un système bibliographique, alors que la fonction de recherche consiste à filtrer la visualisation sur la base de l'entité en fonction du mode choisi. Dans leur approche, la sémantique de PaperVis est construite à partir des citations d'un article et de la catégorisation fondée sur les mots-clés obtenus dans un jeu de documents recherchés (Panchenko *et al.*, 2013). PaperVis contient plusieurs modes de recherche et d'exploration d'articles. Par exemple dans le mode de référence de citation, les utilisateurs spécifient une portée de niveau du réseau de citations comme point de départ pour explorer les articles.

(Schenk *et al.*, 2009) ont proposé SemaPlorer, une application qui permet d'explorer et de visualiser de manière interactive des données hétérogènes et distribuées sur le web. SemaPlorer intègre et exploite différentes ressources telles que DBpedia, GeoNames, WordNet et des fichiers FOAF.

(N. Chen *et al.*, 2018) proposent AnchorViz, une visualisation interactive qui facilite la découverte des erreurs grâce à l'exploration des données sémantiques. En créant des ancres fondées sur des exemples, les utilisateurs créent une topologie pour répartir les données en fonction de leur similarité avec les ancres et examiner

les incohérences entre les points de données liées sémantiquement.

(Dumas *et al.*, 2012) propose l'outil Visual Data Explorer (ViDaX), un outil de visualisation et d'exploration de grands ensembles de données RDF. ViDaX permet l'extraction d'informations à partir de sources de données RDF et offre des fonctionnalités pour l'analyse de diverses caractéristiques de données ainsi que l'exploration de la structure graphique correspondante. ViDaX offre également aux utilisateurs non experts la possibilité d'explorer des données sémantiques fondées sur de puissantes techniques de visualisation et d'interaction automatiques sans aucune programmation de bas niveau .

Avec l'importance croissante de la recherche dans la visualisation d'informations, les visualisations fondées sur les entités gagnent de plus en plus d'importance. Le but de ces visualisations n'est généralement pas de fournir un schéma visuel et une vue d'ensemble des données.

4.2.4 Quelle approche choisir pour le graphe de connaissances ?

Les différentes techniques abordées dans les parties précédentes (visualisation d'information ou de connaissances) dépendent toutes du type de données à visualiser. Ces approches permettent de présenter, souvent de manière interactive, la structure ou la sémantique des informations. En effet, dans notre cas les connaissances sur les interactions entre les facteurs de risque dont on dispose sont modélisées sous forme d'un graphe composé de concepts et de relations. Notre objectif est de faciliter la visualisation et la navigation dans ce graphe. Les graphes RDF et les ontologies sont très puissants, mais peuvent être complexes. Des interfaces utilisateur visuelles intuitives peuvent réduire considérablement la charge cognitive des utilisateurs lorsqu'ils travaillent avec ces structures complexes. La visualisation est une technique prometteuse pour améliorer la perception de la structure par les utilisateurs dans les grands espaces d'information et fournir des moyens de navigation.

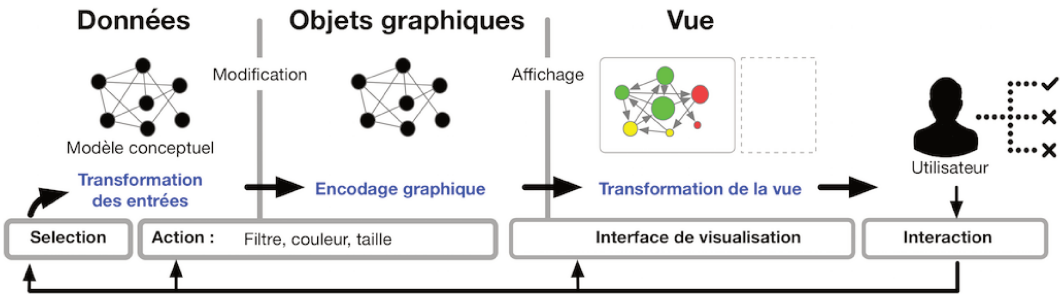


Figure 4.16 – Approche de visualisation dans MCVGraphViz

La démarche de conception que nous suivons correspond à celle présentée dans la figure 4.16, elle est inspirée de celle proposée en visualisation d'information par (Card *et al.*, 1999). Les données sont stockées sous forme d'un modèle conceptuel. L'approche adoptée consiste à charger les données dans un premier module pour construire les variables graphiques. Pour chaque élément du graphe, nous déterminons la façon dont il est dessiné. Enfin, le visuel correspondant est affiché dans l'interface de visualisation. Cette architecture permet à l'utilisateur d'interagir et de manipuler le graphe. Les interactions de l'utilisateur sont finalement retransmises à différents niveaux de cette architecture, en fonction du besoin. Par exemple, lorsque l'utilisateur sélectionne,

zoome ou se déplace dans l'écran, c'est au niveau du rendu visuel que l'action est transmisse.

4.3 Conception de MCVGraphViz

4.3.1 Scénario d'usage

Dans MCVGraphViz l'objectif été de choisir l'approche la plus adaptée pour transférer les connaissances vers l'utilisateur (traité dans le chapitre 5). Sa mise en oeuvre consiste à visualiser les connaissances de manière progressive en proposant des vues complémentaires de granularité plus fine. Les travaux en perception visuelle montrent que l'être humain a une perception d'abord globale d'une scène, avant de porter son attention sur les détails. Ce principe (*overview first, zoom and filter, then details on demand*) est bien connu en visualisation, il a été introduit par (Shneiderman, 1996). Pour cela, de nombreuses approches ont été développées sur ce principe. Par exemple, les approches *overview + détails et focus + context*. Pour une présentation plus complète, l'utilisateur doit donc pouvoir passer aisément d'une vue globale à une vue détaillée en interagissant avec la visualisation. Ce constat nous a conduit à intégrer ce modèle dans l'interface de MCVGraphViz.

Le processus débute avec l'affichage global du modèle conceptuel sous forme d'un graphe décrivant l'ensemble des facteurs de risques et leurs interactions. Les nœuds du graphe correspondent aux facteurs de risques et les interactions entre les facteurs sont représentées par des arcs orientés liant les facteurs. Cependant, la principale difficulté rencontrée dans l'affichage des graphes concerne la surcharge relative aux informations représentées. Nous avons intégré un moteur de recherche dans le graphe pour pallier cette difficultée. Par exemple, imaginons que l'utilisateur veuille identifier les interactions associées à un facteur de risque. Deux stratégies sont possibles : (i) sélectionner un nœud du graphe comme une entrée ; (ii) lancer une recherche dans le graphe. A l'issue de l'une de ces actions, seuls les nœuds connectés au nœud sélectionné seront actifs et colorés suivant leurs catégories, les autres nœuds deviennent inactifs. L'utilisateur peut ainsi rapidement identifier les nœuds associés au nœud sélectionné, et peut également avoir accès à un ensemble de métadonnées sur ce nœud.

4.3.2 Encodage graphique

L'encodage graphique concerne la façon dont les variables vont être visualisées dans l'interface. Comme nous l'avons annoncé précédemment, nous disposons d'un graphe ce qui nous amène à représenter deux variables : des nœuds et des liens. La figure 4.17 présente la liste des variables employées pour traduire celle du graphe.

Les nœuds sont symbolisés par des cercles pleins dont : (i) le diamètre est défini par le nombre de liens associés avec ce dernier. Cela permet rapidement d'identifier un nœud d'intérêt par rapport à l'autre ; (ii) la couleur est définie par la catégorie à laquelle le nœud appartient. En utilisant cette variation du diamètre et de la couleur, il devient très simple de détecter les nœuds les plus importants dans le graphe et s'ils sont connectés à des nœuds de taille plus réduite ou au contraire de taille plus importante. Afin de différencier le nœud correspondant à l'entrée du graphe, une variation de luminosité est intégrée autour du nœud.

Les relations entre les nœuds sont représentées par des lignes dirigées et étiquetées indiquant le sens de la relation. Pour différencier ces liens l'épaisseur est définie par

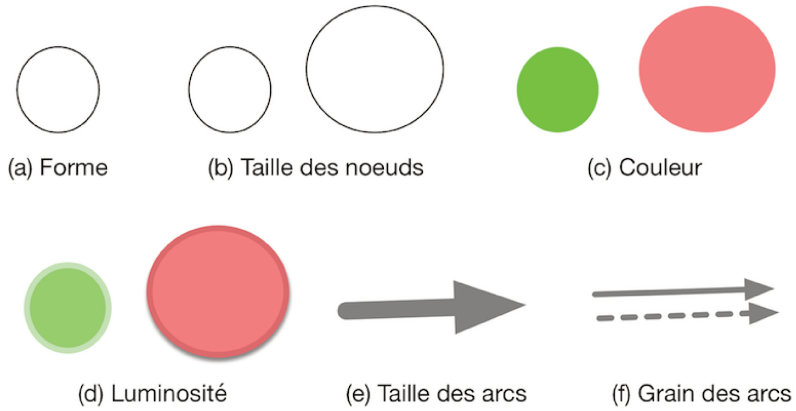


FIGURE 4.17 – Liste des variables graphiques employées dans MCVGraphViz

la probabilité de l'association entre deux nœuds et le grain est défini par la catégorie des deux nœuds associés. Cela permet rapidement d'identifier la relation d'intérêt par rapport à l'autre.

Concernant la clarté du graphe, si cela n'est pas très important lorsque le nombre de nœuds et de liens est faible, cela peut avoir un réel impact lorsque la taille du graphe est importante. En effet, un graphe dense peut conduire à une visualisation non compréhensible à cause du chevauchement des éléments du graphe. Pour résoudre ce problème, nous avons pris en compte une autre variable bien connue en visualisation qui est le positionnement. Afin d'éviter tout croisement inutile entre les éléments du graphe (nœud, relation, étiquette, etc.), chacun est positionné de la manière la mieux adaptée.

4.3.3 Aspect dynamique

4.3.3.1 Interaction avec la visualisation

La plupart des interactions dynamiques commencent par une opération de sélection, dans laquelle un ou plusieurs composants du graphique sont isolés pour certaines actions, tels que la mise en évidence, le masquage, le déplacement ou l'obtention de détails. Par exemple, pour désencombrer un graphique, il est possible de sélectionner un nœud et de désactiver le reste des éléments qui ne sont pas connectés. Ainsi, nous obtenons une visualisation moins chargée, tout en conservant leurs liens. De même, il est possible de sélectionner et de déplacer ou de modifier la forme d'un lien pour éliminer un croisement ou améliorer l'esthétique d'un graphique. La sélection peut impliquer un objet unique, tous les objets d'une région ou distance spécifiée ou un ensemble d'objets satisfaisant à un ensemble de contraintes spécifié par l'utilisateur (par exemple, tous les nœuds directement connectés à un nœud spécifique).

L'un des problèmes les plus importants liés à la sélection d'éléments dans un graphe se produit dans des régions denses du dessin, où les éléments sont si rapprochés qu'une sélection sans ambiguïté est difficile ou impossible. Cette situation conduit à la nécessité d'utiliser d'autres types d'interaction, tels que le zoom ou les techniques de distorsion.

Un autre élément visuel connu est celui du mouvement perceptible par l'utilisateur. Son utilisation est courante dans certaines zones de visualisation pour mettre en évidence les modifications d'un ensemble de données ou d'informations sélectionnées par l'utilisateur (par exemple, clic sur un nœud, survole avec la souris, etc.).

MCVGraphViz est conçu pour visualiser les connaissances contenues dans un modèle conceptuel on permettant son interprétation. Dans ce contexte, l'utilisateur peut avoir besoin d'informations supplémentaires sur certains éléments du graphe (nœud, relation, etc.). Pour conséquent, MCVGraphViz fournit cette fonctionnalité selon le principe de Shneiderman (*overview first, zoom and filter, then details on demand*) (Shneiderman, 1996). L'avantage de l'approche adoptée dans MCVGraphViz est qu'elle permet d'une part de visualiser l'ensemble du graphe et d'autre part d'affiner l'accès aux connaissances représentées dans le modèle (par exemple, un système d'info-bulles).

4.3.3.2 Acquisition de nouvelles connaissances

Nous faisons l'hypothèse que les parcours du graphe reflètent des connaissances implicites réutilisables dans le contexte d'une évaluation du risque. En effet, chaque action réalisée (choix d'un nœud, clic souris sur un nœud, enchaînement de liens entre des nœuds, etc.) exprime une connaissance propre à l'auteur du parcours. Un système de collecte des actions effectuées au cours de la navigation permet de constituer une trace et de la stocker sous la forme d'un chemin (nœud de départ, nœud de destination).

Un système dédié à la gestion de traces permet de collecter des traces, les stocker et les manipuler à l'aide d'opérations génériques appelées transformations qui sont de différents types : filtrage, fusion de traces, etc (Fuchs, 2017). La trace s'enrichit à chaque nouvelle interaction au fur et à mesure que l'activité se déroule, en prenant en compte le contexte de chaque action réalisée. C'est donc un objet dynamique, qui peut être disponible en temps réel (Besnaci *et al.*, 2015) et qui peut être modifié (Champin *et al.*, 2013), requêté ou visualisé.

L'utilisation des traces d'interactions vise le plus souvent : (i) l'analyse *a posteriori* des usages et des activités ; (ii) la qualification des usages (Rossi *et al.*, 2005) ; (iii) l'extraction de motifs fréquents dans l'activité tracée (Georgeon *et al.*, 2006) et étudier les comportements des utilisateurs à partir de ces motifs. MCVGraphViz collecte les actions effectuées au cours de la navigation, constitue une trace et la stoke sous la forme d'un chemin (nœud de départ, nœud de destination).

4.3.4 Implémentation

4.3.4.1 Architecture

MCVGraphViz est construit sur une architecture client-serveur. La visualisation du graphe s'effectue côté client. L'objectif principal est de développer une interface de visualisation dynamique. Il était donc important d'intégrer les variables graphiques listées *supra* et les fonctionnalités interactives à fournir aux utilisateurs. Ainsi, la conception et le développement de l'interface graphique de visualisation du graphe sont entièrement fondés sur SVG. L'avantage présenté par l'utilisation du format vectoriel SVG dans les applications de gestion des connaissances est qu'il offre de nombreuses fonctionnalités interactives aux utilisateurs, conformes aux stratégies ergonomiques de communication homme-machine.

Pour le rendu visuel des variables graphiques, nous avons choisi la librairie graphique D3.js. Contrairement à d'autres bibliothèques, qui produisent un élément graphique sans garder la trace entre la donnée et le graphique, D3.js met en œuvre une approche différente. Il transforme les données en modèle d'objet de document (DOM [6]). Ainsi, les éléments graphiques sont conservés dans un arbre DOM et dessinés sur un canevas.

6. `https://www.w3.org/DOM/`

Par conséquent, les projections sont facilement réversibles puisqu'elles s'appuient sur les éléments du DOM. Par exemple, la mise en évidence d'éléments ou la réaction à des actions de sélection peuvent être gérées de manière uniforme. La figure 4.18 présente l'interaction entre les données, les éléments DOM et les éléments graphiques.

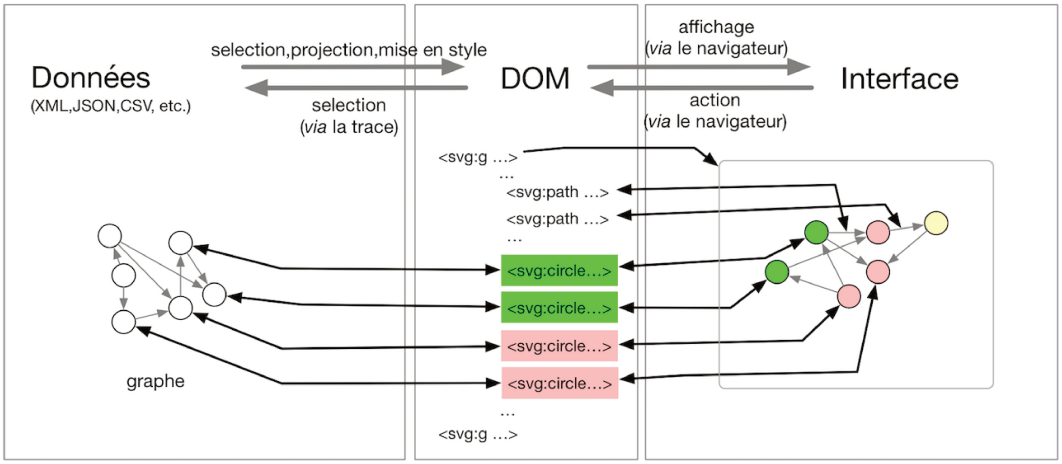


FIGURE 4.18 – Processus d'interaction entre : D3.js, DOM et l'interface de visualisation.

L'utilisateur peut cependant vouloir récupérer des informations sur certains nœuds ou liens du graphe. Par conséquent, MCVGraphViz permet d'afficher ces informations qui sont stockées dans une base de données à l'aide de MySQL et PHP. Les interfaces graphiques nécessaires à l'utilisation des outils d'interaction ont été implémentées en JavaScript, HTML5 et CSS3.

Les solutions adoptées concernant les implémentations réalisées sont temporaires. D'autres technologies seront envisagées par la suite. Le Système de Gestion de Base de Données (SGBD) envisagé sera orienté graphe de la famille NoSQL (par exemple, Neo4j). L'avantage de ce type de système réside dans son approche de description de données et sa très grande lisibilité pour l'utilisateur. De plus, cette nouvelle organisation s'accompagne d'une rapidité de traitement des requêtes beaucoup plus importante que les SGBD classiques. Notons qu'une version en D3.js et Neo4j a été implémentée en parallèle de la version proposée dans ce travail. L'architecture client-serveur, elle sera probablement en Node.js, car elle permet un développement multi-plateformes (mobile, tablette, etc.) tout en restant dans un seul écosystème (sans passer d'un langage de programmation à l'autre).

4.3.4.2 Résultats

Nous allons présenter ici un cas d'utilisation avec un jeu de données issues de l'étude de (MENETON *et al.*, 2016). Ce jeu de données présente les interactions entre les facteurs de risque cardiovasculaires sous forme d'un modèle conceptuel (issue de l'extraction réalisée dans le chapitre précédent). Le nombre de concepts et de relations dans ce modèle conceptuel est respectivement 13 et 58.

Nous pouvons ainsi voir dans la figure 4.19 l'interface de visualisation développée dans MCVGraphViz. Dans cette interface, nous distinguons trois zones :

1. Zone (a) : correspond au graphe avec les facteurs de risque et leurs interactions. Une légende est associée au graphe pour guider l'utilisateur dans la lecture du graphe. Chaque facteur de risque est présenté par un cercle. La couleur des cercles varie suivant la catégorie à laquelle le facteur de risque appartient. De plus, le

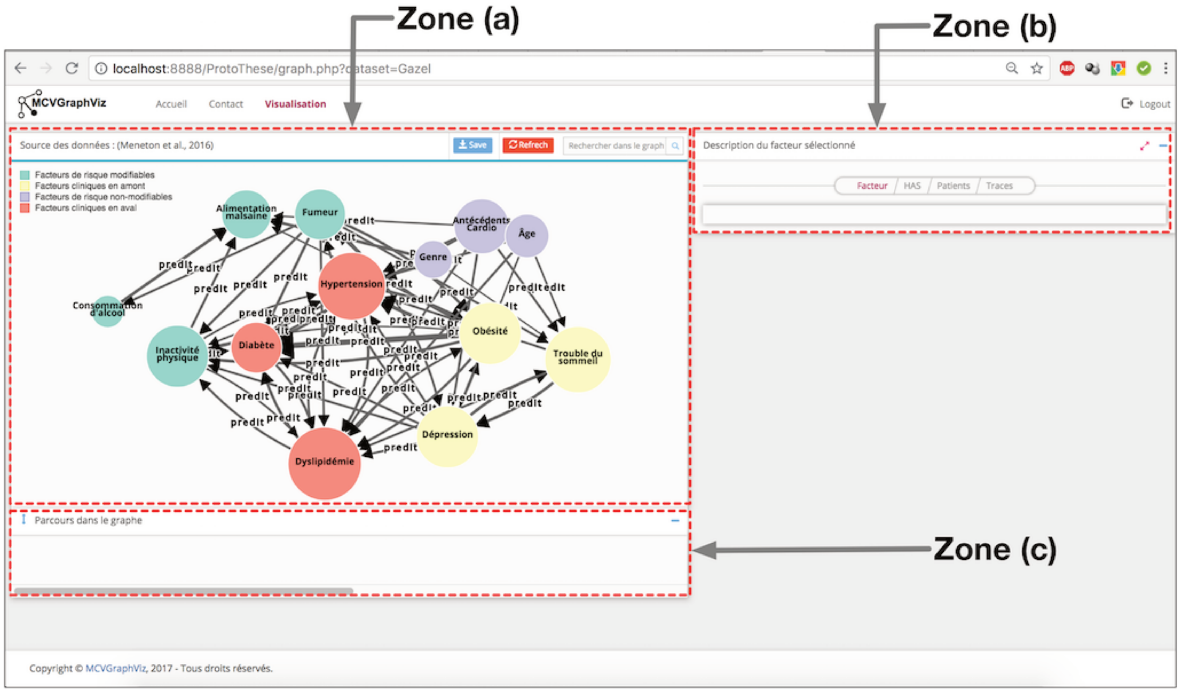
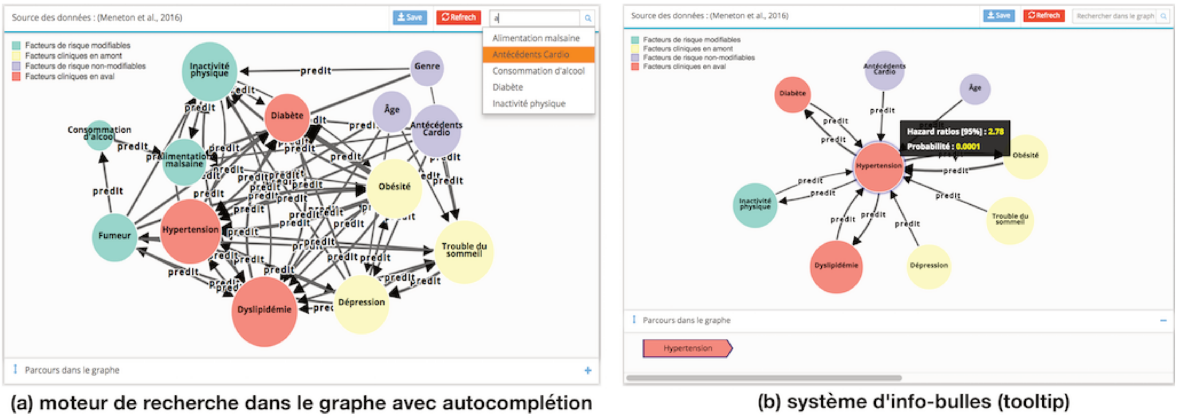


FIGURE 4.19 – Interface de l'outil MCVGraphViz

diamètre des cercles dépend du nombre de liens le composant. En examinant le graphe, on constate un partitionnement des facteurs de risque selon leurs catégories en quatre *clusters*. Les relations entre les facteurs de risque sont orientées et les étiquettes sont positionnées au centre de la relation. Pour éviter que les relations se chevauchent, lorsque le nombre de liens entre deux nœuds dépasse un, ces relations subissent une torsion pour obtenir un des arcs. Afin de ne pas surcharger la densité de l'affichage avec des étiquettes de relations trop élevées, nous avons intégré un système d'info-bulles (voir figure 4.20(b)) pour fournir plus d'informations sur la nature de la relation. Il en va de même pour la lecture dans le graphe, nous avons intégré des effets permettant de désactiver des nœuds avec le survol de la souris pour fournir plus de visibilité à l'utilisateur. Dans le même ordre d'idée, nous avons intégré un moteur de recherche dans le graphe avec autocomplétion (voir figure 4.20(a)) pour guider l'utilisateur dans la sélection des facteurs de risque ;



(a) moteur de recherche dans le graphe avec autocomplétion

(b) système d'info-bulles (tooltip)

FIGURE 4.20 – Exemple d'action intégrée dans MCVGraphViz.

2. Zone (b) : correspond à une zone d'accès à des informations complémentaires lorsqu'un facteur de risque est sélectionné (voir figure 4.21). Parmi ces informations figure sa catégorie, un résumé des interactions le concernant et de recommandation de la haute autorité de santé (HAS [7]), etc. Cette zone contient aussi un formulaire de connexion à une base de données permettant la récupération et le stockage de profil patients. Ainsi, nous pouvons partir du profil du patient comme entrée dans le graphe de visualisation et évaluer ainsi le risque auquel il est exposé ;

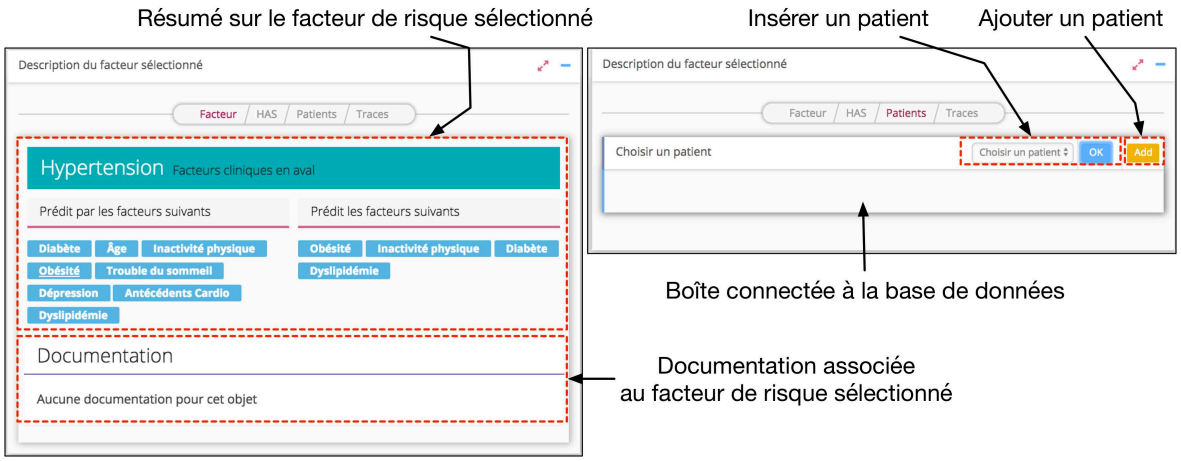


FIGURE 4.21 – Exemple d'informations contenu dans la zone (b).

3. Zone (c) : correspond à l'historique des sous-graphes visités dans le parcours actuel (voir figure 4.22). Chaque nœud visité est ainsi affiché dans l'ordre des clics dans le graphe. Ainsi, l'utilisateur peut se situer dans sa navigation et effectuer des retours arrières si besoin. A l'issue de son parcours, cette navigation est stockée sous forme d'une trace, qui sera exploitée ultérieurement.

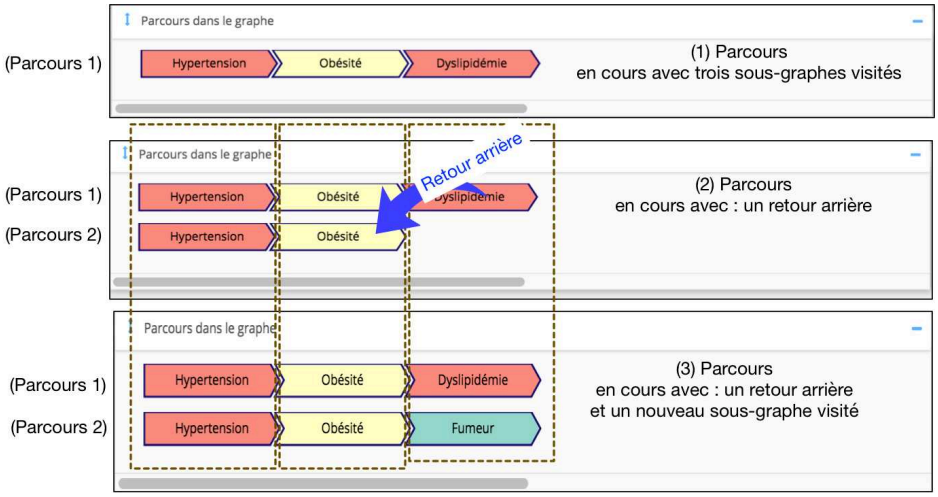


FIGURE 4.22 – Exemple de parcours dans le graphe

Lorsque l'utilisateur effectue des actions dans l'interface, chacunes des zones précédemment décrites s'actualisent en temps réel. Par exemple, si l'utilisateur sélectionne le facteur de risque « Hypertension », le résultat de cette action est présenté dans la

7. https://www.has-sante.fr/portail/

figure 4.23. Dans la zone (a), le sous-graphe associé au facteur de risque « Hypertension » est produit. De la même manière, sur cette nouvelle représentation les interactions associées au facteur de risque « Hypertension » sont visualisées. L'importance du facteur est représentée par son nombre de connexions et sa taille. Cette visualisation fait apparaître que le facteur « Hypertension » prédit les facteurs « Obésité », « Inactivité physique », « Diabète » et « Dyslipidémie » et qu'il n'y a qu'un seul élément appartenant à la catégorie des « Facteurs de risque modifiables ».

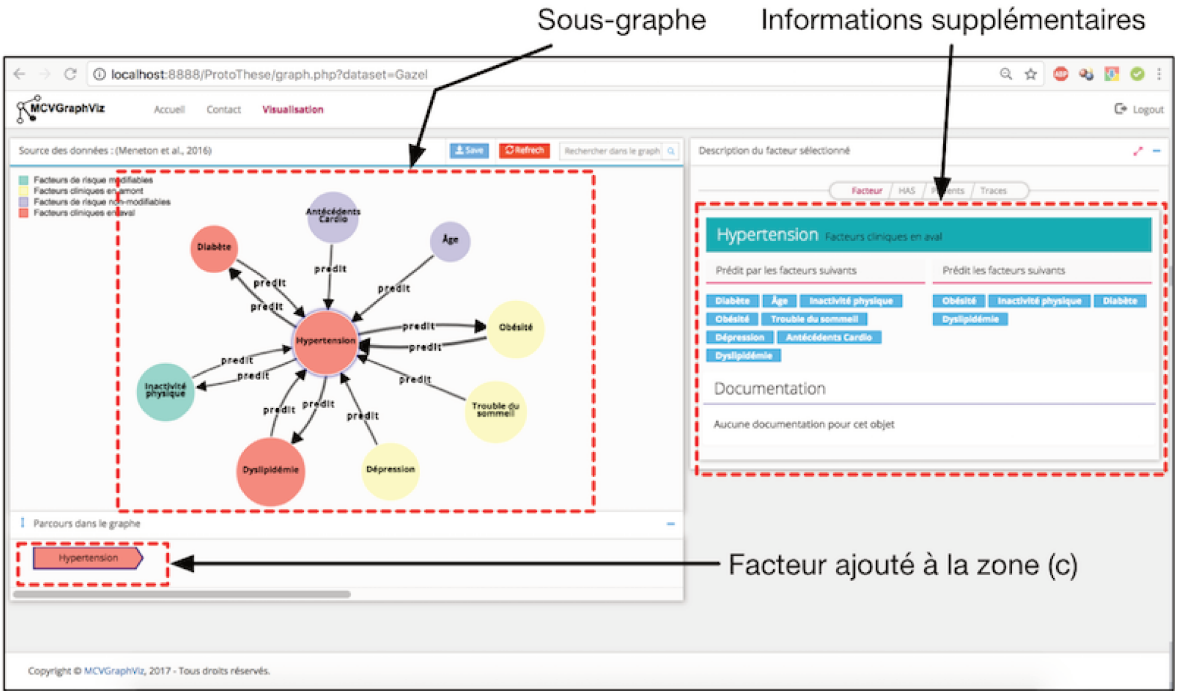


FIGURE 4.23 – Interface de MCVGraphViz actualisée après la sélection d'un facteur de risque

Souvent en visualisation de graphe, il se pose toujours la question du passage à l'échelle lorsqu'il y a beaucoup de nœuds et de liens à visualiser. Dans notre expérimentation cette problématique n'est pas observée car le nombre de nœuds et de liens restent acceptables. Cependant, si le problème se pose, il est possible de le réduire soit : (i) en s'appuyant sur le moteur de recherche pour filtrer les étiquettes des nœuds ; (ii) en couplant le zoom et le déplacement sur des régions du graphe pour diminuer le nombre de nœuds et de relations visibles.

4.4 Conclusion

Nous avons présenté un nouvel outil de visualisation dédié aux connaissances structurées sous forme de graphe. Cet outil permet de visualiser le graphe, et d'interagir avec le modèle conceptuel représentant les connaissances du domaine. Nous avons réutilisé les résultats recensés dans l'état de l'art pour obtenir des visualisations équivalentes aux outils dédiés à la visualisation de données. Nous avons donc conçu et développé une méthode permettant d'accéder de façon progressive au contenu du graphe. Il est possible de visualiser le graphe dans son ensemble ou d'en obtenir une vue partielle en désactivant certains nœuds et arcs. Ce processus permet de se focaliser sur des éléments dans le graphe et de proposes différentes vues choisies, précises et restreintes.

De plus, notre technique permet de collecter et stocker les actions d'interaction réalisées par l'utilisateur sous forme de traces. L'outil MCVGraphViz a évolué au fur et à mesure des versions (voir annexe A) pour répondre à des besoins spécifiques à partir des échanges que nous avons eus avec un médecin. L'évaluation de l'interface de l'outil MCVGraphViz est en cours auprès d'internes et de médecins généralistes. MCVGraphViz sera utilisé en pratique dans le cadre de l'écosystème modulaire qui sera présenté dans le chapitre 5. L'objectif est d'avoir une évaluation précise du risque cardiovasculaire pour pouvoir associer des recommandations adaptées.

Chapitre 5

Systèmes de recommandation

Sommaire

5.1 Introduction **104**
5.2 Les systèmes de recommandation **104**
5.2.1 Définitions 104
5.2.2 Classification des systèmes de recommandation 105
5.3 Systèmes de recommandation en santé **112**
5.3.1 Composantes des systèmes de recommandation en santé 113
5.3.2 Comparaison des systèmes de recommandation en santé 117
5.4 Approche proposée **119**
5.4.1 Utilisateurs du système 119
5.4.2 Méthode de recommandation 120
5.5 Conclusion **126**

Le système de recommandation est le sujet du cinquième chapitre. Nous commençons par nous intéresser aux systèmes de recommandation et leur utilisation dans le domaine de la santé. Nous nous intéressons à la fois aux aspects techniques et à la manière dont les utilisateurs perçoivent un tel système. Puis nous proposons une approche de recommandation modulaire adaptée à l'utilisateur en nous fondant sur leurs besoins et un ensemble de critères identifiés dans la littérature. Après avoir abordé le choix de la modularité, nous présenterons les interfaces implémentant chaque module alliant ainsi l'expertise humaine à l'efficacité informatique pour améliorer la prévention du risque cardiovasculaire.

5.1 Introduction

Depuis leur lancement au début des années 1990, les systèmes de recommandation automatisés ont révolutionné le marketing et plus particulièrement les plateformes e-commerce et des recommandations de contenus (film, musique, etc.) en proposant des recommandations et des prédictions personnalisées sur une large gamme de produits complexes et de grande taille (Konstan et Riedl, 2012).

A l'origine, les systèmes de recommandation sont apparus pour tenter de résoudre les problèmes liés à la surcharge informationnelle [1]. Les systèmes de recommandation sont des systèmes de filtrage d'information traitant le problème de la surcharge d'informations (Konstan et Riedl, 2012). Ils filtrent les fragments d'informations à partir d'une grande quantité d'information générées dynamiquement en fonction des préférences, des intérêts ou du comportement observé de l'utilisateur (Pan et W. Li, 2010). Un système de recommandation possède la capacité de prédire si un utilisateur particulier est susceptible de préférer un item (film, plat, etc.) ou non en fonction de son profil. Les différents domaines où les systèmes de recommandation peuvent être utilisés (Srinivasa *et al.*, 2018) sont associés à la mode, à la cuisine, au jeu, à la musique, au cinéma, etc.

Le principe de la recommandation est la mise en relation de deux entités : des utilisateurs et des items. Les items pertinents sont mis en avant pour les utilisateurs. Le niveau de pertinence est déterminé par le système en fonction d'informations telles que le profil de l'utilisateur, le contexte de consommation, les items disponibles, l'historique des transactions, les rétroactions (*feedbacks*) d'autres utilisateurs sur l'item, etc. L'utilisateur peut alors parcourir les recommandations et fournir un retour implicite ou explicite sur la pertinence de ces items. Par exemple, dans une plateforme de e-learning les rétroactions peuvent être exprimées par des notes ou des avis attribués aux contenus par les apprenants. Toutes les actions et les rétroactions des utilisateurs peuvent être enregistrés dans la base de données du système de recommandations, et par la suite utilisés pour produire de nouvelles recommandations.

5.2 Les systèmes de recommandation

5.2.1 Définitions

Un système de recommandation est défini comme un ensemble de stratégies visant à prendre la décision pour les utilisateurs dans des environnements d'information complexes (Rashid *et al.*, 2002). Le but d'un système de recommandation est de proposer à un utilisateur des items susceptibles de correspondre à son profil. Il permet de réduire de manière considérable le temps de recherche des items par l'utilisateur et aussi de trouver des items liés à ses intérêts et ses préférences (J. B. Schafer *et al.*, 1999). Les systèmes de recommandation ont été appliqués dans de nombreux domaines, tels que les documents (Liu, Lai *et al.*, 2012), les livres (Mooney et Roy, 2000), la musique (Lee *et al.*, 2017), les services adaptés (Liu, K. Chen *et al.*, 2018), etc.

Les termes les plus souvent associés à un système de recommandation sont :

1. La surcharge informationnelle, surinformation ou infobésité (terme francisé d'invention québécoise, issu de l'anglais, information overload) est un concept désignant l'excès d'informations reçues par une personne qu'elle ne peut traiter ou supporter sans porter préjudice à elle-même ou à son activité.

	Tarte aux pommes	Riz au lait	Nougat glacé
Utilisateur 1	4	5	2
Utilisateur 2	4	4	?
Utilisateur 3	2	?	5

TABLE 5.1 – Exemple d'une matrice de notes

— *Utilisateur :* correspond à des personnes utilisant le système et recevant des recommandations. Les utilisateurs fournissent également les notes pour les articles.

— *Item :* correspond à l'objet de la recommandation, il varie selon le domaine (il peut être : un film dans le domaine du cinéma, une recette de cuisine dans le domaine de l'alimentation, etc.).

— *Note :* correspond à une préférence exprimée par l'utilisateur pour un *item* proposé. Cependant, la majorité des systèmes utilisent des évaluations formulées à l'aide d'une échelle allant de 1 à 5, ou bien des évaluations binaires (j'aime/je n'aime pas).

— *Tâche :* correspond à la raison pour laquelle l'utilisateur fait appel au système (par exemple, recherche d'un film).

— *Contexte :* correspond à divers facteurs contextuels influençant l'utilisation et l'évaluation du système de recommandation.

— *Problème de démarrage à froid :* correspond aux problèmes rencontrés par le système de recommandation lorsqu'il est nécessaire de proposer de nouveaux items ou lorsque de nouveaux utilisateurs intègrent le système pour lesquels nous n'avons aucune information explicite ou implicite (GOTARDO *et al.*, 2013).

— *Couverture :* correspond à l'utilisation de tous les items de la base de données pour les recommandations. Par exemple, dans une base de données de 100 films, nous devons recommander au moins une fois chacun des films pour avoir une bonne couverture.

— *Matrice des notes :* correspond au triplet (*utilisateur, item, note*). Lorsque l'utilisateur ne donne pas de notes à la recommandation, on obtient une paire (*utilisateur, item*). Le tableau 5.1 monte un exemple de matrice de notes pour trois utilisateurs et trois recettes de cuisine. Comme nous pouvons le voir, certaines recettes n'ont pas eu de notation (correspond au champ « ? »).

Contrairement aux autres domaines, la recommandation dans le domaine de la santé peut être définie comme il suit : « *A recommendation is a practical suggestion to implement an action aimed at improving the client's functional outcome and quality of life.* » (MARCOPULOS, 2017).

5.2.2 Classification des systèmes de recommandation

Il est possible de classer les systèmes de recommandation de différentes manières. En général, les systèmes de recommandation peuvent être : (i) *non-personnalisés*, ils fournissent aux utilisateurs des suggestions, mais ne prennent pas en considération les préférences des utilisateurs (Par exemple, la liste de chansons qui font le plus de passages radio sur une période donnée (top-50)) ; (ii) *personnalisés*, en utilisant les caractéristiques des utilisateurs, leurs préférences, leurs profils ou bien leurs interactions passées afin de fournir des suggestions de contenu pertinentes. Ces suggestions

peuvent aussi être envisagées comme étant des prédictions sur les interactions futures de ces mêmes utilisateurs. Deux grandes approches de systèmes de recommandation personnalisés existent : les systèmes fondés sur le contenu, les systèmes fondés sur le filtrage collaboratif (Adomavicius et Tuzhilin, 2005). A ces deux types d'approches, (Burke, 2007) s'ajoute trois approches développées pour répondre aux faiblesses de ces dernières : la recommandation fondée sur des données démographiques, la recommandation fondée sur des connaissances et la recommandation fondée sur l'utilité.

Nous présentons par la suite les approches fondées sur le contenu et le filtrage collaboratif, puis les approches hybrides et enfin, les systèmes de recommandation sensibles au contexte. L'objectif est d'introduire ces systèmes de recommandation et leurs évolutions pour répondre et s'adapter à des besoins et des problématiques du monde réel et d'identifier dans quel contexte ces systèmes peuvent être utilisés. Nous avons choisi ces approches, car l'objectif est de construire un écosystème apprenant capable : (i) de recommander aux utilisateurs des contenus pertinents ; (ii) d'exploiter les avis des utilisateurs par rapport à des recommandations ; (iii) de s'adapter au contexte particulier de la santé.

5.2.2.1 Approche fondée sur le contenu

Les méthodes de recommandation fondées sur le contenu utilisent des techniques largement inspirées du domaine de la recherche d'information. L'approche consiste à proposer à l'utilisateur un choix parmi de nouveaux items jugés proches des items qu'il a précédemment appréciés. Dans ce processus de recommandation, nous pouvons distinguer deux ensembles : (i) les profils d'items, constitués à partir des items déjà consultés sur la base des items aimés ou détestés ; (ii) les profils d'utilisateurs, construits à partir des rétroactions sous forme d'un profil « positif » pour les items aimés et « négatif » pour les items détestés. L'efficacité de la recommandation dépend de la qualité du profil construit.

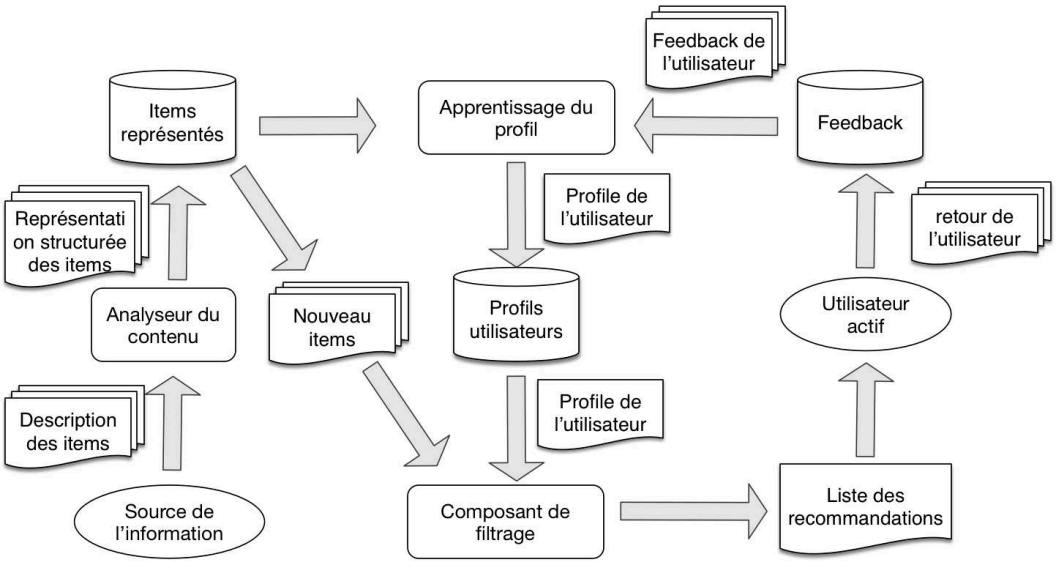


Figure 5.1 – Une architecture haut niveau d'un système de recommandation fondé sur le contenu (Lops *et al.*, 2010)

Pour être pertinent, un système de recommandation doit produire une représentation efficace des items, du profil de l'utilisateur et des stratégies adaptées pour comparer les items du système avec ceux du profil. Parmi les travaux allant dans ce sens, (Lops

et al., 2010) proposent une architecture de haut niveau (voir figure 5.1). Le processus de recommandation se déroule en trois étapes :

Analyseur de contenu : vise à prétraiter les sources d'information non structurées (par exemple, texte des pages web, description d'un produit, etc.) pour extraire des informations pertinentes et les structurer dans un format adapté (par exemple, une page web représentée par un vecteur de mots clés). La précision est dépendante de la nature des contenus : elle est plus élevée pour des contenus textuels et moins élevée pour des contenus vidéo ou des images dont il est difficile d'extraire des attributs.

Apprentissage du profil : utilisé pour collecter les données représentatives des préférences de l'utilisateur et pour généraliser ces données afin de construire le profil utilisateur. Habituellement, la stratégie de généralisation est réalisée à l'aide de techniques d'apprentissages automatiques (par exemple les arbres de décisions, les réseaux de neurones, etc.) qui permettent d'inférer un modèle des intérêts des utilisateurs à partir des items appréciés ou non.

Composant de filtrage : utilisé pour suggérer des éléments pertinents appariant la représentation du profil avec celle des items pertinents à recommander. La pertinence de l'item est calculée en utilisant des métriques de similarité entre l'item considéré et le profil de l'utilisateur. L'évaluation de la pertinence peut être binaire ou continue (liste classée d'éléments potentiellement intéressants).

Dans cette architecture pour chaque utilisateur actif, un profil est créé et mis à jour en construisant des rétroactions à partir des réactions aux items recueillis ou à partir de ses intérêts initialement définis. Ces centres d'intérêts sont exploités au cours du processus d'apprentissage du modèle pour prédire la pertinence *a priori* d'un item que l'utilisateur n'a pas encore sélectionné.

La plupart des systèmes de recommandation fondés sur le contenu reposent sur le modèle vectoriel VSM (Vector Space Model) avec la pondération classique TF-IDF (Term Frequency-Inverse Document Frequency) pour la représentation spatiale de documents texte. Dans ce modèle, chaque document (désignant un item) est décrit par un vecteur de dimension n, où une dimension correspond à un terme de l'ensemble du vocabulaire d'une collection de documents. Formellement, tout document est représenté par des vecteurs de poids sur les termes, où chaque poids indique le degré d'association entre le document et le terme. Soit $D = \{d_1, d_2, ..., d_N\}$ un ensemble de documents ou corpus, et $T = \{t_1, t_2, ..., d_n\}$ le dictionnaire, c'est-à-dire l'ensemble des termes contenus dans le corpus. T est généralement obtenu en appliquant des opérations de traitement du langage naturel, comme l'atomisation (tokenization), l'élimination des mots vides de sens, et la troncature (stemming) (Baeza-Yates et Ribeiro-Neto, 1999). Chaque document d_j est représenté par un vecteur dans un espace vectoriel à n dimensions, tel que $d_j = \{w_{1j}, w_{2j}, ..., w_{nj}\}$, où w_{kj} est le poids du terme tk dans le document d_j.

La représentation des documents dans le VSM soulève deux problèmes : pondérer les termes et mesurer la similarité des vecteurs d'entités. Le schéma de pondération des termes les plus communément utilisés est la pondération TF-IDF :

$$TFIDF(t_k, d_j) = \underbrace{TF(T_k, d_j)}_{TF} . \underbrace{log\frac{N}{n_k}}_{IDF} \tag{5.1}$$

où N désigne le nombre de documents dans le corpus, et n_k représente le nombre de

documents de la collection dans lesquels le terme t_k apparaît au moins une fois, avec :

$$TF(t_k, d_j) = \frac{f_{k,j}}{max_z f_{z,j}} \in [0, 1] \tag{5.2}$$

où $f_{k,j}$ représente le nombre d'occurrences du terme t_k dans le document d_j, et $max_z f_{z,j}$ est le maximum des fréquences $f_{z,j}$ des termes t_z apparaissant dans le document d_j. Afin que tous les poids appartiennent à l'intervalle [0, 1] et que tous les documents soient représentés par des vecteurs de même longueur, les poids obtenus par la fonction TFIDF sont généralement normalisés en utilisant la normalisation cosinus :

$$w_{k,j} = \frac{TFIDF(t_k,d_j)}{\sqrt{\sum_{s=1}^{|T|} TFIDF(t_s,d_j)^2}} \tag{5.3}$$

Après cette étape de pondération des termes et de normalisation, il faut définir une mesure de similarité des vecteurs caractéristiques. Cette mesure de similarité est requise pour déterminer la proximité entre deux documents. Il existe de nombreuses mesures de similarité, mais la mesure la plus largement utilisée dans la littérature est la similarité cosinus :

$$sim(d_i, d_j) = \frac{\sum_k w_{ki} \cdot w_{kj}}{\sqrt{\sum_k w_{ki}^2} \cdot \sqrt{\sum_k w_{kj}^2}} \tag{5.4}$$

De la même manière, les profils utilisateurs sont représentés par des vecteurs de termes pondérés. La prédiction de l'intérêt d'un utilisateur pour un item donné est ensuite obtenue par un calcul de similarité entre le vecteur profil de l'utilisateur et le vecteur de contenu en utilisant la similarité cosinus (voir formule 5.4).

Les systèmes de recommandation fondés sur les vecteurs mots-clefs ont été développés dans divers domaines d'application, tels que les actualités, la musique, le commerce électronique, les films, etc. Chaque domaine présente différents problèmes qui requièrent différentes solutions : dans le domaine des recommandateurs web, le système Letizia (Lieberman, 1995), dans le secteur de l'actualité (Ahn *et al.*, 2007), dans le domaine collaboratif avec (Diederich et Iofciu, 2006), etc.

Selon (Lops *et al.*, 2010), les approches fondées sur le contenu présentent plusieurs avantages et inconvénients :

Indépendance à l'utilisateur : Les recommandations fondées sur le contenu exploitent uniquement les évaluations fournies par l'utilisateur actif pour créer son propre profil, contrairement aux approches de filtrage collaboratif.

Transparence : des explications sur le fonctionnement du système de recommandation peuvent être fournies en listant explicitement les fonctionnalités de contenu ou les descriptions qui ont provoqué l'apparition d'un élément dans la liste des recommandations.

Nouveaux items : Les recommandations fondées sur le contenu peuvent recommander des éléments qui ne sont pas encore évalués par un utilisateur. Par conséquent, ils ne souffrent pas du problème du premier avis, affectant les recommandateurs collaboratifs qui reposent uniquement sur les préférences des utilisateurs pour formuler des recommandations.

Les systèmes fondés sur le contenu présentent plusieurs inconvénients :

Analyse de contenu limitée : une limite naturelle de la recommandation fondée sur le contenu est la nécessité de disposer d'une représentation variée et riche du contenu des items, ce qui n'est pas toujours le cas. La précision des recommandations est liée à la quantité d'information dont dispose le système pour discriminer les items appréciés de ceux non appréciés par l'utilisateur.

Sur-spécialisation (Over-specialization) : la recommandation fondée sur le contenu propose des éléments dont les scores sont élevés par rapport au profil de l'utilisateur. C'est pourquoi l'utilisateur recevra parfois des items similaires à ceux déjà notés.

Intégration des nouveaux utilisateurs non immédiat : lorsque peu d'évaluations sont disponibles, pour un utilisateur, le système ne pourra pas fournir de recommandations fiables. Ce problème est connu dans la littérature sous le nom du problème de démarrage à froid pour les utilisateurs (*user cold start*).

D'autres formes de recommandation fondée sur le contenu existent, par exemple les **recommandations fondées sur la sémantique**. Le principe consiste à exploiter les technologies du web sémantique afin de collecter le maximum d'informations sur un utilisateur pour pouvoir ensuite lui recommander des items. L'analyse sémantique (à l'aide d'ontologie) permet d'apprendre des profils plus précis contenant des références à des concepts définis dans des bases de connaissances externes. La principale motivation de cette approche est de fournir un système de recommandation avec les connaissances sur les pratiques culturelles et linguistiques caractérisant la capacité d'interpréter des documents en langage naturel et de raisonner sur leur contenu.

SiteIF est un agent personnel pour un site web d'information bilingue qui apprend les intérêts de l'utilisateur à partir des pages demandées (Magnini et Strapparava, 2001). Il adopte une représentation documentaire fondée sur le sens afin de construire un modèle représentant les centres d'intérêts des utilisateurs. La ressource externe impliquée dans le processus de représentation est MultiWordNet [2].

SEWeP (amélioration sémantique pour la personnalisation web) (Eirinaki *et al.*, 2003) est un système de personnalisation web qui utilise à la fois les historiques d'utilisation et la sémantique du contenu d'un site web afin de le personnaliser.

Recommender Informated (Aciar *et al.*, 2007) utilise des revues de produits de consommation pour faire des recommandations. Le système convertit les opinions des consommateurs en une forme structurée en utilisant une ontologie de traduction, qui est exploitée comme une forme de représentation et de partage des connaissances.

En ce qui concerne les méthodes de recommandation fondées sur le contenu, toutes les études intègrent des connaissances linguistiques ou propres à un domaine ou les deux. Des méthodes de filtrage fondées sur le contenu ont fourni des résultats meilleurs et plus précis par rapport aux méthodes traditionnelles fondées sur le contenu. Cela encourage les chercheurs à concevoir de nouvelles méthodes de filtrage qui formalisent et contextualisent les centres d'intérêts des utilisateurs en exploitant des ressources externes telles des thésaurus ou des ontologies (Lops *et al.*, 2010).

5.2.2.2 Approche fondée sur le filtrage collaboratif

Le filtrage collaboratif est la deuxième grande famille de systèmes de recommandation. Il s'appuie sur le partage d'opinions entre les utilisateurs afin de les regrouper. L'idée suppose que si des utilisateurs ont les mêmes préférences sur un ensemble d'items, alors ils auront probablement les mêmes préférences sur un autre ensemble d'items qu'ils n'ont pas encore notés. Contrairement à l'approche fondée sur le contenu, ici il n'est pas nécessaire d'analyser le contenu des items. En effet, il suffit de connaître la proximité des utilisateurs en fonction de leur profil.

Dans les approches proposées nous pouvons distinguer généralement deux sous-familles principales : (i) les approches *fondées sur des heuristiques* ; (ii) les approches

2. `http://multiwordnet.fbk.eu`

fondées sur les voisins qui exploitent les notes des utilisateurs stockés pour faire de la prédiction. Dans ces approches, les algorithmes utilisés commencent par la construction d'une matrice de notes pour réduire la complexité des calculs ou de traiter le problème des notes manquantes.

Filtrage collaboratif fondé sur les voisins. L'approche de filtrage collaboratif fondée sur les voisins ou appelée filtrage *k plus proche voisins* (ou *k-Nearest Neighbours-kNN*) consiste à mettre en place des algorithmes qui exploitent la totalité de la matrice des notes des utilisateurs pour établir la recommandation (Nakamura et Abe, 1998). Ces approches sont regroupées en deux familles : (i) fondées sur les utilisateurs (*user-user collaborative filtering*), consistant à prédire l'appréciation des utilisateurs par rapport à un item en s'appuyant sur les notes de ses voisins (utilisateur avec qui il partage les mêmes préférences) (Pujahari et Padmanabhan, 2015 ; Resnick *et al.*, 1994) ; (ii) fondés sur les items (*item-item collaborative filtering*), consistant à estimer l'appréciation d'un item i candidat par l'utilisateur à partir des notes attribuées pour les items voisins de i (Linden *et al.*, 2003 ; Miranda et Jorge, 2009 ; Sarwar *et al.*, 2001).

Filtrage collaboratif fondé sur les méthodes de réduction de la dimension. Dans les modèles de filtrage collaboratif, les utilisateurs fournissent une évaluation sur de multiples aspects d'un item dans de nouvelles dimensions, augmentant ainsi la taille de la matrice de notation. Les méthodes de réduction de la dimension sont donc nécessaires pour résoudre ces problèmes afin de réduire la dimension de la matrice de notation des items pour améliorer la précision des prévisions et l'efficacité du système de recommandation pour cette réduction. L'approche utilisée consiste à projeter les items (et/ou les utilisateurs) dans une dimension réduite définie par des variables latentes afin de traiter le problème de la sensibilité aux données manquantes (Ekstrand *et al.*, 2011) et le problème du passage à l'échelle (B. Schafer *et al.*, 2007). Cette disposition dans un espace plus dense et plus réduit permet d'identifier de nouvelles relations entre des paires d'utilisateurs même s'ils n'ont aucun item en commun.

Les techniques de réduction de dimension sont devenues un premier choix pour la mise en oeuvre des systèmes de recommandation (Koren et R. Bell, 2010), surtout après le challenge lancé par Netflix (Koren et R. Bell, 2015). L'analyse de ces approches a démontré la supériorité en termes de précision des approches appliquant des techniques de réduction de la dimension par rapport aux algorithmes de filtrage collaboratif fondé sur les voisins. Les méthodes de factorisation de matrice sont utilisées pour l'extraction des variables latentes telles que l'Analyse en Composante Principale (ACP) (Goldberg *et al.*, 2001) ou la Décomposition en Valeur Singulière (SVD en anglais *Singular Value Decomposition*) (Koren, 2008). Elles ont été essentiellement utilisées, soit pour réduire la dimension de la matrice des notes, soit pour réduire la dimension de la matrice de similarité (Desrosiers et Karypis, 2010). Pour plus de détails sur les récents travaux appliquant la factorisation de matrice dans les algorithmes de recommandations collaboratives, le lecteur peut se référer à (Koren et R. Bell, 2015).

Filtrage collaboratif fondé sur les modèles probabilistes. Le principe de ces approches consiste à modéliser le comportement des utilisateurs par un modèle probabiliste pour pouvoir prédire ses comportements futurs. L'idée sur laquelle reposent ces algorithmes consiste à calculer la probabilité que l'utilisateur u attribue la note v à l'item i connaissant ses notes antérieures. La prédiction pred

(u, i) correspond, soit à la note ayant la plus grande probabilité, soit à la note espérée (B. Schafer *et al.*, 2007).

(B. M. Kim et Q. Li, 2004) proposent une approche de modèle probabiliste fondée sur les items afin de résoudre les problèmes de biais lié à l'utilisateur, d'association non transitive et d'élément nouveaux dans l'espoir d'obtenir de meilleures performances. Dans ce modèle probabiliste, les items sont classés en groupes et des prévisions sont faites pour les utilisateurs en tenant compte de la distribution gaussienne des évaluations des utilisateurs. (Campos *et al.*, 2008) proposent de classer automatiquement une liste de nouveaux éléments à un utilisateur en fonction des habitudes de votes antérieures d'autres utilisateurs ayant des goûts similaires. L'approche combine des réseaux bayésiens qui permettent une représentation intuitive des mécanismes régissant les relations entre les utilisateurs, et la théorie des ensembles flous, qui permet de représenter l'ambiguïté ou le flou dans la description des évaluations, améliore la précision du système.

5.2.2.3 Approche hybride

Les approches hybrides sont définies par (Burke, 2002) comme une combinaison de deux ou plusieurs approches de recommandation. Ces combinaisons permettent de bénéficier des avantages des approches utilisées, de pallier leurs inconvénients et de proposer des recommandations plus pertinentes. Elles sont actuellement les plus représentées dans la littérature, notamment parce qu'elles sont jugées comme étant les plus efficaces (Schein *et al.*, 2002). Il existe plusieurs manières de faire de l'hybridation et aucun consensus n'a été défini par la communauté des chercheurs. Toutefois, (Burke, 2002) a identifié sept manières différentes de faire l'hybridation (Burke, 2007). Par exemple, l'hybridation à bascule, dans laquelle le système passe d'une approche à l'autre en fonction de certains critères, l'hybridation mixte qui consiste à présenter à l'utilisateur des recommandations issues de plusieurs approches de recommandations, etc.

5.2.2.4 Approche de recommandation sensible au contexte

L'importance des informations contextuelles a été reconnue comme un paramètre important qui joue un rôle dans la pertinence des recommandations. Cependant, les travaux de recherche sur les systèmes de recommandation ont assez peu exploité les informations contextuelles.

Les systèmes de recommandation sensible au contexte consistent à intégrer des données liées au contexte de l'utilisateur telles que sa position, l'heure, les éléments disponibles par rapport à sa position. Ces systèmes sont à la base des systèmes de recommandations classiques (filtrage fondé sur le contenu ou filtrage collaboratif) qui intègrent de nouvelles composantes pour l'acquisition et le traitement du contexte. Le filtrage consiste ensuite à analyser le contexte courant de l'utilisateur et à sélectionner, parmi les contenus disponibles, ceux qui correspondent le mieux à la situation décrite.

Pour intégrer le contexte, (Karatzoglou *et al.*, 2010) proposent des méthodes de factorisation tensorielle. Pour ces méthodes, en plus des deux premières dimensions traditionnellement utilisées pour les items et les utilisateurs, chaque type de contexte est considéré comme une nouvelle dimension. La note n'est plus considérée comme une fonction avec les deux paramètres item et utilisateur, mais une fonction ayant comme paramètres l'item, l'utilisateur et les aspects du contexte.

Avec la mise en place d'un système de recommandation se pose la question de comment identifier l'approche la mieux adaptée à son contexte. Comme nous l'avons

précédemment présenté, il existe plusieurs approches avec des avantages et des limites. Les approches les plus récentes tentent d'apporter des solutions aux problèmes identifiés en combinant plusieurs approches dites hybrides. Les approches hybrides sont des approches qui combinent deux ou plusieurs approches de recommandation. Ces combinaisons permettent de bénéficier des avantages des approches utilisées, de pallier leurs inconvénients et de proposer des recommandations plus pertinentes.

5.3 Systèmes de recommandation en santé

L'intérêt des professionnels de santé envers les systèmes de recommandation est due aux possibilités offertes, pour : (i) récupérer des informations supplémentaires pour un cas particulier, tel que des directives cliniques ou des articles de recherche connexes ; (ii) fournir aux utilisateurs finaux des contenus de haute qualité, fondés sur des preuves et liés à la santé.

La plupart des systèmes de recommandation identifiés dans la littérature proposent des actions qui se concentrent sur les utilisateurs (patients) pour produire des recommandations pertinentes. Ces approches consistent à fournir des informations fiables aux utilisateurs, comme dans les travaux de (Wiesner et Pfeifer, 2014), les recommandations de changement de mode de vie (Farrell *et al.*, 2012) et l'amélioration de la sécurité des patients (Roitman *et al.*, 2010). Par exemple, les recommandations de changement de mode de vie sont formulées à l'utilisateur sous forme de suggestion relative à un comportement à adopter pour améliorer l'alimentation (Pinxteren *et al.*, 2011 ; Ueda, Asanuma *et al.*, 2014), l'activité physique (Ali *et al.*, 2018) ou le sommeil (Datko *et al.*, 2016), etc.

L'aspect éducatif est un point essentiel dans la santé, l'utilisation des ordinateurs a commencé il y a plus de dix ans, principalement pour adapter les ressources en matière d'éducation sanitaire. Ces systèmes ne tirent pas parti de la quantité croissante de ressources éducatives disponibles sur internet. L'une des raisons est due à la quantité énorme d'informations à traiter. Ce problème de la surcharge d'information a été abordé pendant de nombreuses années dans le domaine des systèmes de recommandation (Fernandez-Luque *et al.*, 2009). De plus, les systèmes de recommandation en santé sont présentés comme des outils complémentaires dans les processus de prise de décision dans les services de santé. Dans une brève revue sur les systèmes de recommandation en santé, (Sezgin et Ozkan, 2013) ont souligné leur importance croissante. En effet, ces systèmes présentent un potentiel d'amélioration de la facilité d'utilisation des dispositifs de soins de santé en réduisant la surcharge d'information produite par les appareils et les logiciels médicaux et améliorent ainsi leur acceptation.

Les questions liées à la santé dans les systèmes de recommandation ont constitué un sujet de recherche ces dernières années. A titre d'exemple, nous pouvons citer le premier atelier sur la gestion de la santé avec les systèmes de recommandation (Elsweiler, Ludwig *et al.*, 2016). L'objectif de cet atelier était de rassembler des chercheurs et des praticiens de divers domaines (santé, bien-être, aide à la décision et changement de comportement) pour partager leurs expériences.

5.3.1 Composantes des systèmes de recommandation en santé

5.3.1.1 Comprendre le domaine

Pour tout domaine d'application la mise en place d'un système de recommandation nécessite d'apporter des réponses à certaines questions. La première question concerne la nécessité de comprendre quel est le type de la recommandation. Dans ce contexte, les composantes identifiées dans la littérature pour le domaine de la santé peuvent être classées comme suit :

1. **Information nutritionnelle :** de nombreux travaux de recherche ont mis en évidence le lien entre le comportement alimentaire et certaines maladies. Par exemple, le diabète, le cancer et l'hypertension ont pour origine certaines habitudes alimentaires malsaines (Mulrooney et J. Bell, 2016). Le problème peut être résolu en mettant en place des systèmes de recommandation. L'approche consiste à fournir des suggestions dans le but d'améliorer le profil nutritionnel d'une personne. Par exemple, pour aider à réduire un risque d'obésité ou pour prévenir certaines maladies liées à l'alimentation (Elsweiler, Harvey *et al.*, 2015 ; Fernandez-Luque *et al.*, 2009). Les recommandations peuvent se présenter sous diverses formes : (i) proposition d'aliments de substitution (Achananuparp et Weber, 2016 ; Freyne et Berkovsky, 2010) ; (ii) proposition de repas variés (Ge *et al.*, 2015) ; (iii) proposition de compléments alimentaires.
   D'autres paramètres impliqués dans l'acceptation de la recommandation doivent être pris en compte lors du processus de recommandation. Par exemple, la perception sensorielle des recettes (Harvey *et al.*, 2012), le contexte social des individus (Berkovsky et Freyne, 2010 ; Ueda, Morishita *et al.*, 2016), etc.
   (Geleijnse *et al.*, 2011) proposent un système de recommandation de recettes personnalisées qui aide les utilisateurs à faire des choix de repas sains en fonction de ses choix antérieurs. Pour stimuler l'adoption d'un mode de vie plus sain, un mécanisme de définition d'objectifs est appliqué en combinaison avec des suggestions de recettes personnalisées.
   (Nazer *et al.*, 2014) ont conçu un questionnaire par lequel un client répond à diverses questions concernant son mode de vie, ses préférences alimentaires, son apport en compléments et ses habitudes alimentaires. Une fois ces données collectées, l'utilisateur aura accès à une planification pour une semaine avec des repas distincts pour chaque jour.
   (Helmy *et al.*, 2015) proposent un système permettant de récupérer des informations personnalisées pour répondre aux questions des utilisateurs sur les domaines de l'alimentation, de la santé et de la nutrition. Par exemple, « La pomme peut-elle être bénéfique pour les personnes atteintes de maladies cardiaques ? ».
   (Yang, C. K. Hsieh *et al.*, 2017) proposent Yum-me, un système de recommandation de repas personnalisés, conçu pour répondre aux attentes nutritionnelles, à la prise en compte de restrictions alimentaires et aux préférences alimentaires des individus. Yum-me permet une procédure de profilage des préférences alimentaires simple et précise *via* une interface utilisateur fondée sur un quiz visuel.
   (Bianchini *et al.*, 2017) présentent le système de recommandation d'aliments PREFer, capable de fournir aux utilisateurs des menus personnalisés et sains, en tenant compte des préférences à court et à long terme et des prescriptions médicales de l'utilisateur.
   (Subramaniyaswamy *et al.*, 2018) proposent ProTrip RS un système fondé sur

les motivations et les préférences des utilisateurs pour proposer des suggestions. Les données prises en compte pour les recommandations incluent la séquence de déplacement, le contexte, etc.

2. **Exercice physique :** consiste à fournir des recommandations sur les activités physiques à effectuer. L'objectif est de proposer des activités intéressantes et motivantes et correspondant également aux besoins des utilisateurs. Un conseiller peut également inclure des données d'emplacement et des données météorologiques pour trouver des activités optimales pour le contexte des utilisateurs.

   (Izumi *et al.*, 2006) ont utilisé le formalisme OWL pour modéliser les zones d'activité physique et individuelle pour fournir des suggestions de bien-être et d'exercice physique. Les données des espaces individuels et des espaces d'exercice sont regroupées et représentées avec une structure ontologique.

   (Erzsébet *et al.*, 2010) ont proposé une approche traduisant la phase de construction de planification alimentaire et de l'activité physique entre experts et patients. Un modèle hiérarchique formel représentant le problème de planification des menus ou des exercices est utilisé.

   (Chiang *et al.*, 2014) proposent de planifier un programme quotidien et personnalisé sous forme d'activité physique pour contrôler l'apport calorique. À l'aide des mises à jour quotidiennes et des commentaires d'experts, les résultats de la recommandation sont ajustés.

3. **Diagnostic :** consiste à fournir à un médecin ou à une infirmière des recommandations sur les diagnostics probables pour un patient. Cette approche de recommandation peut être très similaire aux approches de raisonnement à partir de cas.

   (Kung *et al.*, 2011) proposent un nouveau système de recommandation de services alimentaires personnalisés intelligents (IPFSRS) qui comprend une couche web de capteurs vitaux (VSWL), une couche web médical sémantique (SMWL) et une couche de présentation des services médicaux (MSPL). Les capteurs vitaux du VSWL peuvent transférer les données cliniques des utilisateurs sur la base de l'activation du capteur (SWE). Le SMWL utilise le raisonnement fondé sur des règles (RBR) et les ontologies de domaine fondées sur le web sémantique (SW) pour déterminer l'état de santé d'un utilisateur en fonction des données de cet utilisateur provenant de VSWL.

4. **Thérapie :** consiste à fournir des recommandations sur les différentes possibilités pour un traitement. Un système de recommandation peut s'adresser à la fois au patient et au professionnel pour déterminer une thérapie adaptée au patient. La thérapie pourrait inclure diverses propriétés du patient et visualiser différents critères de résultat. Les systèmes de recommandation peuvent créer une santé personnalisée uniquement à partir d'une analyse de données.

L'identification des utilisateurs du domaine étudié. Généralement, le système est conçu pour un utilisateur final, qui peut être en bonne santé (dans le cas de la prévention) ou déjà patient (amélioration de la prise en charge). En outre, les systèmes de recommandation en matière de santé peuvent faire intervenir des professionnels de la santé tels que les médecins, les infirmières, etc.

Le contexte d'utilisation du système contient à la fois la définition d'objectifs multifactoriels et des facteurs contextuels qui influencent la manière dont les items sont recommandés et comment ils doivent être présentés. L'objectif est de ne pas ce contenter de recommander l'option la plus saine sans prendre en compte, différents critères

propres à un domaine qui jouent un rôle dans la réussite de la recommandation. Par exemple, ce qui est sain pour un patient peut être dangereux pour un autre. Il est nécessaire d'inclure la réduction des effets secondaires qui peuvent exister (recommander d'arrêter de fumer pour une personne obèse peut déclencher une dépression).

L'accès aux données est un paramètre important. Contrairement à d'autres domaines de recommandation (par exemple, les films, les articles, etc.) qui disposent d'un accès public facile aux données, l'accès dans le domaine de la santé est restreint. Cela est dû parfois à la nature complexe des données de santé qui sont souvent non structurées, incomplètes, non normalisées et proviennent de diverses sources. Beaucoup de ces données sont disponibles au format papier dans les dossiers patients qui sont souvent numérisés par la suite. De plus, le type de données peut être décrit de différentes manières et à différents niveaux. Par exemple dans le domaine de la nutrition bien que nous ayons de grandes bases de données nutritionnelles, il est difficile de comprendre quels sont les aliments substituablent adaptés au goût de l'utilisateur. Différentes marques de produits alimentaires similaires rendent également les recommandations plus difficiles à faire.

Enfin, les utilisateurs doivent également être identifiés et pris en compte dans le système. Par exemple, dans le cas d'une recommandation alimentaire la nourriture est souvent consommée en groupe, de sorte que les préférences individuelles de plusieurs utilisateurs influencent la possibilité de choix. Des choix ou des recommandations très spécifiques (par exemple, faible teneur en glucides, faible en gras, sans gluten) peuvent être contradictoires.

5.3.1.2 Type d'algorithme utilisé

De nombreuses approches sont proposées dans la littérature. Elles combinent des techniques utilisant les filtrages fondés sur le contenu et le filtrage démographique en s'appuyant sur des modèles ontologiques pour améliorer les résultats de la recommandation.

L'objectif est de renforcer la précision des recommandations en intégrant différentes exigences (par exemple, conditions sanitaires, besoins nutritionnels, préférences pour les aliments et exercices) dans la procédure de recommandation. Le but est de considérer plus de principes et de contraintes dans la procédure de proposition pour améliorerer les suggestions. Par exemple, recommander des aliments et des activités physiques répondant aux besoins de santé et aux besoins nutritionnels des individus consiste à prendre en compte leurs préférences. Dans cette situation, le système de recommandation requiert des efforts supplémentaires de la part des individus car ils doivent déclarer la consommation quotidienne d'aliments de manière cohérente, ce qui peut les amener à se détourner de la recommandation.

5.3.1.3 Aspect éducatif des systèmes de recommandation

La diversité d'approches montre que le profilage des utilisateurs, en particulier pour des secteurs et des domaines tels que la recommandation concernant l'alimentation et la santé, est principalement abordée de manière ponctuelle, sans nécessairement viser à fournir des effets éducatifs aux utilisateurs (Bianchini *et al.*, 2017).

(Abidi et H. Chen, 2006) présentent un système de planification longitudinal des soins aux patients qui génère automatiquement et de manière proactive des plans de soins de santé adaptés aux patients. L'objectif est d'aider les médecins à déterminer le

meilleur discours sur les soins cliniques en fonction : (i) du profil médical actuel du patient ; (ii) des dernières connaissances médicales ; (iii) des parcours cliniques spécifiques à l'institution et des programmes de formation personnalisés.

(Netisopakul *et al.*, 2010) proposent un système de recommandation utilisant une ontologie. Les principales composantes du système sont les profils personnels des utilisateurs, la base de données sur les aliments et la nutrition et la base de connaissances. L'ontologie alimentaire et nutritionnelle développée est utilisée *via* des règles pour fournir des recommandations alimentaires fondées sur les besoins nutritionnels de l'utilisateur ainsi que sur les préférences alimentaires.

5.3.1.4 Principes d'évaluation expérimentale

L'évaluation est un paramètre très important lors de la mise en place d'une démarche de recommandation. Elle consiste à évaluer plusieurs points :

Confiance. L'évaluation de la confiance est un point très important (O'Donovan et Smyth, 2005). Cela est d'autant plus important lorsqu'il s'agit de système de recommandation en matière de santé, car ils doivent être utilisés pour fournir aux utilisateurs finaux des informations plus proactives et personnalisées concernant leur santé. Cependant, l'évaluation de la confiance reste une question de recherche ouverte. Par exemple, comment évaluer la confidentialité et l'intimité dans l'utilisation de la technologie médicale. La diversité des utilisateurs joue un rôle, l'accent étant mis sur le sexe et l'âge (Ziefle *et al.*, 2011).

(Rupp *et al.*, 2016) ont proposé une approche capable d'évaluer la confiance et la motivation en tant que facteurs prédictifs de l'utilisation d'un pister de fitness. Ils ont développé des échelles de confiance et de motivation spécifiques aux appareils de fitness portables afin de déterminer dans quelle mesure ces facteurs peuvent être liés à la motivation des individus de continuer à utiliser. La conclusion de ce travail est que la confiance technologique et la motivation sont étroitement liées l'une à l'autre.

Satisfaction. La question est maintenant de trouver les algorithmes les plus appropriés pour répondre aux besoins des utilisateurs. Jusqu'à présent, les recherches menées ont porté sur l'amélioration de la précision des systèmes de recommandation.

(Wu *et al.*, 2012) proposent un système de recommandation qui dépasse les critères d'exactitude conventionnels et prend en compte d'autres critères tels que la couverture, la diversité, la sérénité, l'extensibilité, l'adaptabilité, le risque, la nouveauté, etc.

(Herlocker *et al.*, 2004) suggèrent d'examiner les principales décisions relatives à l'évaluation des systèmes de recommandation de filtrage collaboratif : les tâches utilisateur évaluées, les types d'analyses et de jeux de données utilisés, la qualité de la prédiction, l'évaluation des attributs de prédiction autres que la qualité, l'évaluation fondée sur l'utilisateur du système dans son ensemble. Dans le cas des systèmes de recommandation de santé, cette prédiction est particulière, car il existe différents groupes d'utilisateurs pertinents. Les différences d'expertise, les connaissances générales, mais aussi les tâches doivent être comprises pour créer des systèmes de recommandation adaptés aux praticiens de la santé, aux médecins cliniciens, aux chercheurs biomédicaux, aux soignants et aux patients.

Impact réel sur la santé. Même lorsque les utilisateurs respectent sur le long terme les recommandations adaptées à leur profil, la question posée est de savoir si les

changements de comportement ou de traitement suivi entraînent les changements de santé souhaités. La question est de déterminer quel est le paramètre de santé évalué, et surtout comment l'évaluer pour atteindre les objectifs fixés.

Proposer un mauvais film peut vous coûter 90 minutes de votre vie, choisir une mauvaise thérapie pourrait réduire la qualité de vie pendant de nombreuses années. Une mauvaise option dans les premières recommandations pourrait avoir des résultats inacceptables. Le concepteur d'un système de recommandation de santé doit faire preuve de prudence et agir de manière responsable à la fois pour générer des recommandations et pour les communiquer. Dans le cas du renoncement au tabac (Hors-Fraile *et al.*, 2016), le système a plus de difficulté à mesurer son efficacité, car les utilisateurs peuvent ignorer le fait de déclarer avoir fumé une cigarette en raison de la désirabilité sociale.

Avant qu'une approche de recommandation puisse être mise en oeuvre en santé, il faut s'assurer que ces systèmes soient suffisamment fiables (O'Donovan et Smyth, 2005). Les systèmes accessibles au public sur internet, tels que les recommandations collaboratives, présentent un risque pour la sécurité, car les utilisateurs finaux ordinaires ne peuvent pas être distingués des diffuseurs d'informations malveillants. Au-delà de ces risques techniques, de telles attaques peuvent entraîner une dégradation continue de la confiance dans l'objectivité et la précision d'un tel système.

Enfin, si les objectifs de la recommandation sont à long terme alors le système doit prendre en compte les changements du comportement dans l'évaluation. Sans cette précaution, cela peut conduire à des recommandations à court terme ce qui n'est pas compatible avec les objectifs fixés.

5.3.2 Comparaison des systèmes de recommandation en santé

Dans le domaine de la santé, les systèmes de recommandations visent à assister les experts à l'aide des algorithmes, alliant ainsi l'expertise humaine à l'efficacité informatique pour améliorer les soins ou la prévention qui joue un rôle très important. Pour mettre en place un système de recommandation efficace, il est nécessaire d'avoir une approche interdisciplinaire lors de la conception. L'ensemble des critères (voir la section 5.3.1) cité précédemment doivent être pris en compte et le traitement inadéquat de l'un de ces aspects garantit l'échec du système de recommandation et la perte de confiance dans les systèmes de recommandation pour la santé en général.

Pour comparer les différents systèmes de recommandation, nous nous sommes appuyés sur les critères, identifiés dans la littérature. Le tableau 5.2 présente quelques systèmes comparés selon les critères précédemment énoncés. Il est intéressant de noter que, d'un point de vue global, les systèmes sont appliqués pour la recommandation alimentataire et l'activité physique. On remarque également que l'aspect éducatif est une fonction attendue d'un système de recommandation, mais qui n'est pas remplie par la plupart des systèmes et que l'implémentation de cette dernière est observée dans les systèmes qui s'appuie sur les ontologies pour le filtrage ou la recommandation. On peut toutefois regretter que certains systèmes souffrent du caractère expérimental. Or, pour s'assurer que les approches proposées satisfassent les futurs utilisateurs, il nous semble important de les faire évaluer.

De ce bilan comparatif est né le besoin d'un système de recommandation en santé. En s'appuyant sur les standards du web sémantique, le but de ce système est de fournir un environnement intelligent pour aider à améliorer la prévention du risque cardiovas-

Système de recommandation	Algorithme de filtrage	Algorithme de recommandation	Expérimentation	Aspect éducatif
(Abidi et H. Chen, 2006)	Contenu	Contenu	Oui	Non
(Oh *et al.*, 2009)	Contexte	Les données de l'environnement	Oui	Oui
(Netisopakul *et al.*, 2010)	Ontologie	Information non accessible	Non	Oui
(Freyne et Berkovsky, 2010)	Contenu, filtrage collaboratif et hybride	Plus proches voisins	Oui	Non
(Geleijnse *et al.*, 2011)	Les choix passés	Information non accessible	Oui	Non
(Teng *et al.*, 2012)	Collaboratif	SVM	Oui	Non
(Bundasak et Chinnasarn, 2013)	Contenu	Vecteur de décision	Oui	Non
(El-Dosuky *et al.*, 2012)	Ontologie et les heuristiques	TF-IDF avec la mesure de similarité	Oui	Non
(Tumnark *et al.*, 2013)	Ontologie	Information non accessible	Non	Oui (athlètes)
(Lin *et al.*, 2014)	Contenu	Contenu d'une matrice	Oui	Non
(Nazer *et al.*, 2014)	Sémantique	Information non accessible	Oui	Non
(Nikulin, 2014)	Collaboratif	Hybride	Non	Non
(Ge *et al.*, 2015)	Collaboratif	Plus proches voisins	Non	Non
(Ahire et Khanuja, 2015)	Sémantique fondée sur l'ontologie	Hybride	Oui	Oui
(Bianchini *et al.*, 2017)	Sémantique fondée sur l'ontologie	Les concepts	Oui	Oui
(Yang, C.-K. Hsieh *et al.*, 2017)	Quiz visuel	Information non accessible	Oui	Non
(Subramaniyaswamy *et al.*, 2018)	Sémantique fondée sur l'ontologie	Les concepts avec filtrage adaptés	Oui	Oui

Table 5.2 – Comparaison de quelques systèmes de recommandation

culaire. Pour les problématiques identifiées dans les sections précédentes, nous avons essayé d'apporter des solutions adaptées en les intégrants dans ce système.

5.4 Approche proposée

Le cadre que nous proposons dans ce travail peut être envisagé comme une synthèse des critères identifiés dans la littérature qui aide à se faire une idée globale des contraintes. Un système de recommandation dans le cadre d'une prévention est souvent jugé en fonction des critères énoncés dans la section 5.3.1 (et de bien d'autres) et le traitement inadéquat de l'un de ces aspects peut conduire à l'échec du système de recommandation et la perte de confiance dans les systèmes de recommandation pour la santé en général. Il ne traite d'aucun aspect du droit, de la politique et de la médecine directement. Les recommandations proposées pour la prévention du risque cardiovasculaire sont issues d'une approche interdisciplinaire. Afin de concevoir notre approche, nous avons recherché des solutions pour plusieurs problèmes identifiés dans la section 5.3.1.

Pour résoudre ces problèmes, nous avons opté pour une approche modulaire. L'idée est de mettre à disposition des utilisateurs (que nous traitons dans la section 5.4.1) un ensemble d'outils leur permettant d'interagir avec le système pour aboutir à un compromis entre l'acceptabilité de la recommandation et les préférences personnelles. L'approche choisie intègre les experts, alliant ainsi l'expertise humaine à l'efficacité informatique pour améliorer la prévention du risque cardiovasculaire auxquel les utilisateurs sont exposés. Nous pouvons ainsi être certains d'atteindre des objectifs en travaillant sur la collaboration entre les experts et les utilisateurs. Dans notre approche, nous avons commencé par identifier les utilisateurs du système. Ensuite, nous avons traité la recommandation en répondant à trois questions : (i) quel est le type de la recommandation ? ; (ii) quels sont les algorithmes à utiliser pour faire la recommandation ? ; (iii) quel est le support de la recommandation ?

5.4.1 Utilisateurs du système

Identifier les utilisateurs du système revient à imaginer les scénarios d'utilisation, ce qui donne une idée sur les besoins de chacun. Nous avons besoin d'un cadre pour nous aider à concevoir un système de recommandation de santé en collaboration avec les utilisateurs finaux. L'objectif de la conception d'un système de recommandation pour la santé doit toujours intégrer les utilisateurs finaux. Une évaluation réalisée avec la HAS sur l'outil MCVGraphViz (présenté dans le chapitre précédent) nous a permis d'identifier les utilisateurs potentiels de notre système ainsi que les scénarios associés.

1. Utilisateur simple : correspond à une personne qui cherche à avoir une hygiène de vie saine en suivant des suggestions adaptées à son profil.
2. Patient : L'accès patient correspond à un patient qui doit respecter certaines recommandations dans le but de diminuer un risque ou d'améliorer sont état de santé. L'accès de ce patient consiste : (i) à suivre les recommandations de la plateforme ; (ii) à se documenter sur les risques auxquels il est exposé, par conseil de son médecin traitant.
3. Professionnel de santé : consiste à assister les professionnels de la santé dans le cadre d'une démarche d'un médecin ayant accès au profil de ses patients pour suivre leur évolution suite aux recommandations mises en place pour la prise

de décision sur la stratégie à adopter. Il représente le médecin non seulement en tant que consommateur d'informations, mais également en tant que personne capable de manipuler de manière interactive le contenu et les outils. Le médecin en tant qu'expert doit s'assurer que les connaissances des experts sont intégrées dans le processus de recommandation, en trouvant des modèles et en fournissant des connaissances supplémentaires pour améliorer la recommandation. Ainsi, les connaissances spécialisées du médecin sont enrichies d'informations supplémentaires et d'un savoir-faire expert.

4. Patient et professionnel de santé : correspond à un accès du médecin traitant en présence du patient pour prendre une décision construite sur la stratégie à mettre en place pour diminuer son risque. Ainsi, le système de recommandation peut intégrer ces décisions dans le processus de recommandation. Par exemple, le patient préfère diminuer la fréquence de consommation de tabac plutôt que d'arrêter complètement.

5.4.2 Méthode de recommandation

5.4.2.1 Type de recommandation

Dans le cas de l'amélioration de la prise en charge du patient, les recommandations de pratique clinique et de santé publique sont mises en place par la HAS [3] comme des propositions développées méthodiquement pour aider le praticien et le patient à rechercher les soins les plus appropriés pour des caractéristiques cliniques données. Leur but est d'améliorer la qualité des soins et de réduire les coûts en s'appuyant sur des recommandations élaborées par des organismes certifiés tels que la HAS ou les sociétés savantes.

Le suivi d'habitudes alimentaires adéquates, la pratique d'une activité physique suffisante et le maintien d'une corpulence normale sont nécessaires pour maintenir un bon état de santé. Ces habitudes de vie permettent également de réduire l'incidence des maladies chroniques.

Au cours de la dernière décennie, les études épidémiologiques retrouvant une association entre la nutrition et les maladies chroniques se sont principalement focalisées, pour la plus grande majorité d'entre elles, sur un seul aliment voire nutriment. Un intérêt croissant a été porté ces dernières années à l'alimentation dans son ensemble grâce au développement des scores alimentaires.

Les recommandations du PNNS [4] sont traduites sous la forme de repères de consommation (et non sous la forme d'une pyramide alimentaire [5]) et de façon suffisamment généraliste pour être perçues comme atteignables (voir figure 5.2). Pour soutenir ce programme, différents guides destinés au grand public ont été diffusés largement.

Cependant, les recommandations générales ne sont pas suffisantes (Bodenheimer, 2002 ;Williamson, 2009), surtout dans le cadre des recommandations alimentaires. Compte tenu des apports alimentaires de nos jours (avec des valeurs caloriques élevées), nous pouvons constater qu'il existait une mesure plus importante du développement des problèmes liés au bien-être en raison des modes de vie changeants et à forte intensité (pas le temps pour cuisiner, restauration rapide, etc.). Un moyen d'améliorer les

3. `https://www.has-sante.fr/`

4. `http://www.mangerbouger.fr/`

5. la pyramide alimentaire est née en 1992 à l'initiative du ministère de l'agriculture américain. Son objet était de fournir au grand public une base qui permettrait de visualiser la constitution de repas équilibrés sur une journée.

FIGURE 5.2 – Conseils essentiels du Programme national nutrition-santé sur l'alimentation et l'activité physique (source : `http://eduscol.education.fr/cid47662/decouvrir-alimentation.html`)

habitudes alimentaires consiste à fournir des informations nutritionnelles et des propositions alimentaires appropriées. Dans cette situation, les cadres de recommandations alimentaires sont recherchés, car ils ont pour but d'aider les individus à adopter un comportement sain pour réduire un risque ou améliorer l'état de santé. Cependant, le système ne doit pas prendre seulement en compte les préférences des clients pour les aliments et les nutriments, mais doit tenir compte des problèmes médicaux, des nécessités nutritives et des pratiques alimentaires antérieures.

Les cigarettes saines n'existent pas, par conséquent les messages de promotion de la santé liés au tabagisme sont simples et clairs. Les messages liés à l'alimentation sont eux plus complexes, car de nombreux facteurs influent sur les choix alimentaires, notamment l'emballage, la publicité, le marketing, la disponibilité, le prix, la taille des portions et l'environnement. Ces facteurs peuvent être à l'origine de certaines décisions conduisant à des comportements alimentaires malsains. Les décisions alimentaires malsaines sont la cause de mauvaises habitudes alimentaires chez les personnes qui développent des maladies (par exemple, l'obésité, le diabète, etc.) considérées comme facteurs de risque cardiovasculaire. L'amélioration de ces décisions alimentaires apportera une amélioration des habitudes et pourrait être bénéfique pour réduire le risque cardiovasculaire. Il y a plusieurs façons d'influencer le choix des aliments et d'améliorer les habitudes alimentaires, mais toutes ne sont pas aussi efficaces.

Un système de recommandation doit prendre en compte les activités spécifiques de la semaine (événements prévus) qui influencent le temps disponible pour cuisiner. En outre, tous les facteurs deviennent moins importants lorsque l'on mange chez un ami.

Les besoins diffèrent d'un utilisateur à l'autre, certains aimeraient une fonctionnalité de planification (d'une semaine), d'autres ne l'utiliseront pas. Certains trouveront par exemple qu'il serait utile de saisir les ingrédients avec lesquels ils veulent cuisiner. Ainsi, le système de recommandation ne doit pas inclure trop d'ingrédients difficiles à obtenir. Il serait également intéressant de prendre en compte la saison.

Le contexte familial est un facteur important, car les individus ont tendance à préparer un repas pour l'ensemble de la famille, même si, parfois, la taille des portions doit légèrement être différente pour répondre à une contrainte (par exemple, moins de glucides pour la personne atteinte de diabète). Les planifications hebdomadaires de menus, souvent les utilisateurs ne souhaitent pas manger la même recette pendant deux jours consécutifs. Pour certains, manger la même recette deux fois par semaine est un choix (par exemple, s'il y a suffisamment de restes), pour d'autre il devrait y avoir au moins une semaine entre les deux. De plus, lorsqu'on examine des recommandations alimentaires ou nutritionnelles, il est nécessaire d'incorporer les pratiques culturelles.

Bien que nous ayons de grandes bases de données nutritionnelles, il est difficile de comprendre quels aliments peuvent servir de substituts avec un goût semblable. Différentes marques de produits alimentaires similaires rendent également les recommandations plus difficiles à faire.

L'objectif de cette démarche est de proposer une architecture modulaire permettant de définir complètement les services qui vont composer l'écosystème apprenant et décomposer la complexité d'un système pour mieux le contrôler. Chacun des modules est indépendant et sa mise en œuvre réalise un service. Il y a autant de modules que de services identifiés. Les différents services ont été conçus indépendamment, mais fonctionnent ensemble comme un tout homogène. Dans ce contexte, nous avons opté pour le développement de deux modules. Un module permettant une qualification nutritionnelle des recettes de cuisine et un module pour la recommandation d'activité physique.

Recommandation de recettes. L'objectif du module est de fournir de proposer des recettes de cuisine à l'utilisateur adapté à son profil. Les besoins exacts en matière de nutrition sont ainsi déterminés et adaptés en fonction du profil de l'utilisateur. Ainsi, nous pouvons voir la consommation exacte en matière d'énergie, de protéines, etc. De plus, cela nous permettra d'avoir un suivi dans le temps de l'évolution du comportement alimentaire de l'utilisateur. En adoptant cette approche, cela nous permettra de parcourir des collections de recettes de manière significative et inspirante sans que le système ne suggère activement des recettes. Cette approche laisse le choix à l'utilisateur d'inclure ses propres recettes de cuisine et de mettre l'accent sur la suggestion de recettes répondant à ses préférences. Dans le prochain chapitre, nous décrirons comment nous avons automatisé la qualification nutritionnelle des recettes de cuisine et la traduction des valeurs nutritionnelles sous forme visuelle.

Recommandation d'activité physique. L'objectif de ce module est d'améliorer l'endurance cardio-respiratoire, l'état musculaire et osseux, réduire le risque cardiovasculaire et d'autres maladies non transmissibles. Les recommandations d'activité physique sont celles proposées par des organismes certifiés tels OMS, HAS, PNNS, etc. L'utilisateur pourra ainsi, choisir des activités adaptées à son profil pour atteindre des objectifs fixés. L'activité physique englobe notamment les loisirs, les déplacements (par exemple la marche ou le vélo), les activités professionnelles, les tâches ménagères, les activités ludiques, les sports ou l'exercice planifié, dans le contexte quotidien, familial ou communautaire.

5.4.2.2 Algorithmes de recommandation

La méthode de filtrage utilisée est le filtrage hybride, un mélange de filtrage collaboratif, de filtrage fondé sur le contenu et de filtrage fondé sur la connaissance qui offre les avantages de chacun de ces systèmes.

Dans le filtrage collaboratif, les utilisateurs cibles sont recommandés en fonction de la similarité avec les autres utilisateurs ayant des préférences similaires. Dans le filtrage fondé sur le contenu, différentes décisions optimistes sont mises en contraste et les décisions déjà notées par l'utilisateur et les meilleures décisions de coordination sont prescrites. Le filtrage fondé sur les connaissances repose sur les inférences des préférences de l'utilisateur, sur les connaissances spécifiques à un domaine ou sur la manière dont les éléments répondent aux préférences des utilisateurs. Par exemple, quels aliments peuvent servir de substituts. Seuls les utilisateurs enregistrés sont autorisés à consulter, à obtenir des recommandations et à évaluer les emplacements pour que des utilisateurs similaires puissent bénéficier des meilleurs choix.

5.4.2.3 Support de recommandation

Pour favoriser l'adoption du système de recommandation, nous avons intégré l'approche de recommandation dans l'interface MCVGraphViz décrite dans le chapitre 4 avec un accès propre à chaque utilisateur. Cette intégration a pour but d'enrichir MCVGraphViz avec de nouveaux modules et fournir un ensemble d'outils destiné à la prévention du risque cardiovasculaire. En terme de besoin, le système doit permettre aux utilisateurs de personnaliser leur propre système de recommandation et de communiquer les besoins directement dans l'interface. En outre, des visualisations efficaces doivent être utilisées lors de l'affichage des données de santé. La visualisation des données doit répondre à l'objectif du système de recommandation et tenir compte de l'utilisateur final et de ses intentions. Une méthodologie pour garantir cela est la méthodologie de l'étude de conception (Sedlmair *et al.*, 2012), qui s'est avérée efficace même dans la conception du système de recommandation visuel (Valdez *et al.*, 2015). La création d'une visualisation dans un outil de recommandation encourage les utilisateurs à explorer les options et à expliquer les recommandations individuelles (Parra *et al.*, 2014).

Nous avons opté pour une approche, elle a été construite suivant les principes de visualisation traités dans le chapitre 4. Elle est dévissée en trois zones (voir figure 5.3) :

— La zone (A) affiche un résumé des objectifs fixés pour l'utilisateur, actuellement pour l'alimentation et l'activité physique. Ces objectifs sont définis à partir du profil de l'utilisateur et des recommandations de la HAS, PNNS, etc. Pour l'alimentation par exemple, l'objectif peut correspondre à une valeur énergétique qu'il ne faut pas dépasser dans sa planification de menu pour la journée. Pour l'activité physique par exemple, c'est une durée de 30 minutes. A chaque fois que l'utilisateur interagit avec le système, les informations dans cette zone sont actualisées pour guider l'utilisateur dans ses choix.

— La zone (B) aide l'utilisateur à planifier les repas de la journée et les activités physiques pour atteindre les objectifs fixés dans la zone (A). Elle permet de planifier quatre repas dans la journée (petit déjeuner, déjeuner, goûter, et dîner) et les activités physiques à réaliser. La zone (B) est décomposée en deux zones : (B1) destinée a une navigation parmi une liste de recettes adaptées (à partir de

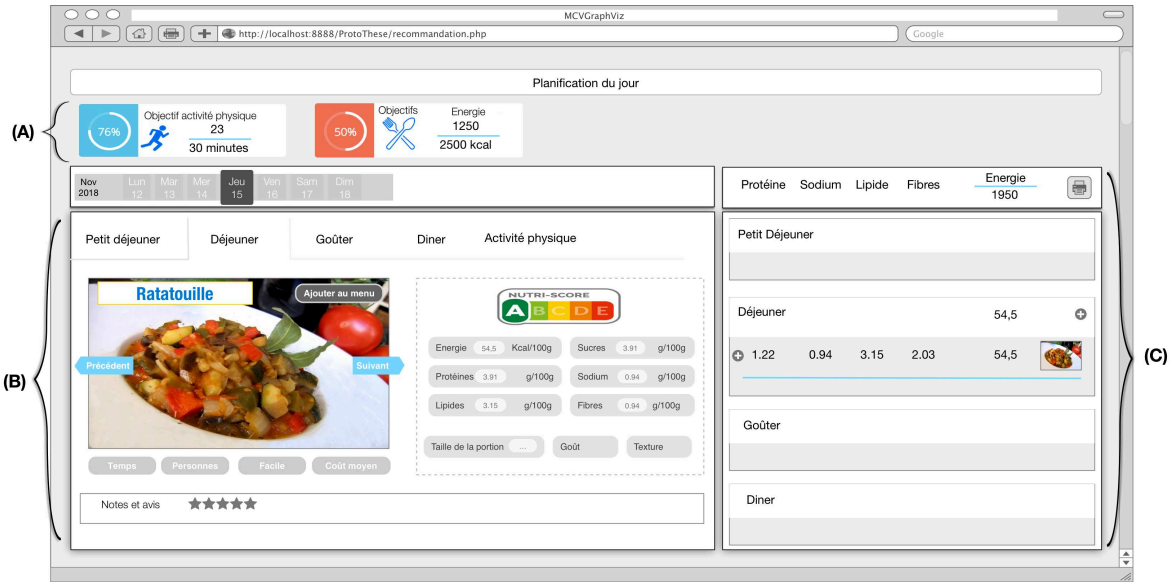


FIGURE 5.3 – Intégration des recommandations dans l'outil MCVGraphViz : accès utilisateur simple ou patient

certaines informations : risque cardiovasculaire, préférences, etc.) à l'utilisateur, de telle sorte qu'il puisse construire lui-même ses menus de la journée. Chaque recette proposée est qualifiée d'un point de vue nutritionnel (cette partie est traitée dans le chapitre 6). Ainsi, elle est décrite par ses valeurs nutritionnelles (énergie, sucres, protéines, etc.), son Nutri-Score, la taille des portions, certaines informations sensorielles (par exemple, le goût, la texture, etc.) et certaines informations comme le temps de préparation, le nombre de personnes, les notes et avis des autres utilisateurs, etc. De plus, l'utilisateur peut ajouter ses propres recettes à la planification grâce à un outil de qualification nutritionnelle automatique (voir figure 5.4), il suffit de saisir la liste des ingrédients et la préparation pour pouvoir réaliser la qualification nutritionnelle.

— La zone (C) à droite de l'écran, affiche un résumé des recettes ou des aliments choisis pour chaque repas de la journée. Chaque repas contient une liste de recettes ou aliments sélectionnées par l'utilisateur à partir de la zone (B). De plus, chaque recette ou aliment est décrit par ses valeurs nutritionnelles. L'utilisateur pourra modifier chaque repas en ajoutant une recette, un aliment ou supprimer ceux déjà choisis. L'utilisateur peut aussi exprimer ses préférences pour des recettes en donnant une note.

Concernant l'interface d'accès des médecins (voir figure 5.5), nous avons intégré un module dans MCVGraphViz permettant au médecin d'accéder au profil du patient et des recommandations associées à ce profil, car elle optimise l'affichage de l'ensemble des connaissances nécessaires au médecin lors de l'évaluation du risque cardiovasculaire. Le médecin sélectionne le profil du patient. Il accède ensuite à son risque cardiovasculaire, sa biologie, son dossier administratif, et les recommandations générées à partir des arbres d'évaluation du risque de la HAS. Ainsi, il peut naviguer visuellement et interagir avec l'interface pour accéder à des informations nécessaires pour la prise de décision sur la stratégie à adoptée. Il peut par exemple voir la consommation lipidique sur une semaine, la fréquence d'activité physique, etc. De cette façon, il peut comparer les recommandations du patient par rapport à celles de la HAS.

Les interfaces de recommandation proposées ont été conçues pour optimiser l'affi-

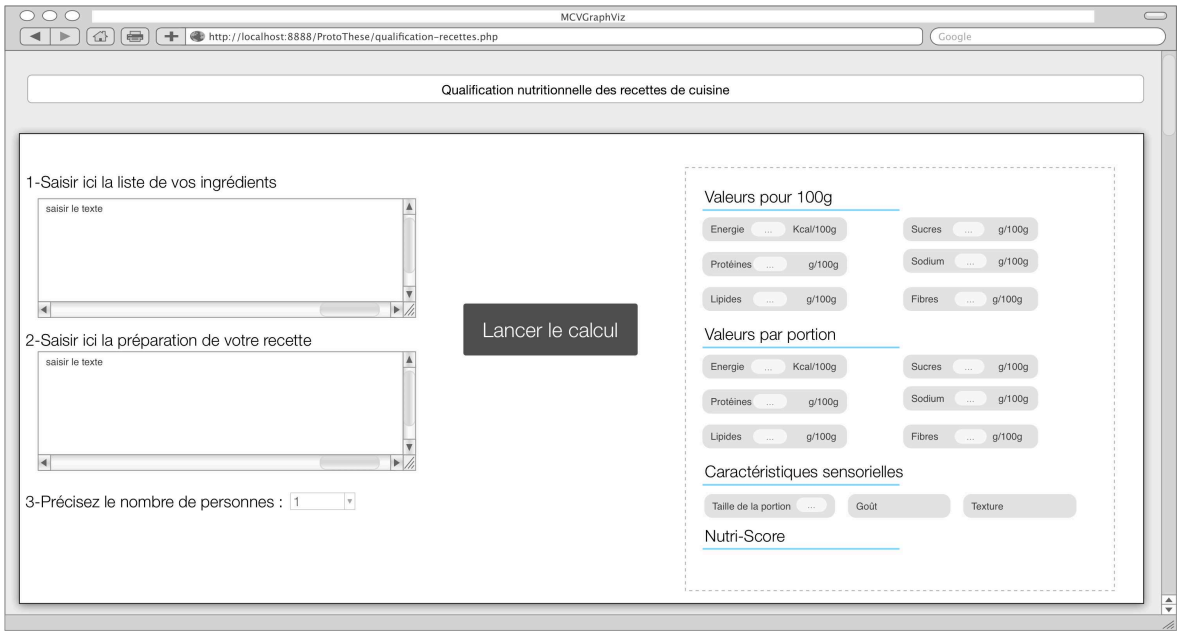


FIGURE 5.4 – Outil de qualification nutritionnelle des recettes de cuisine

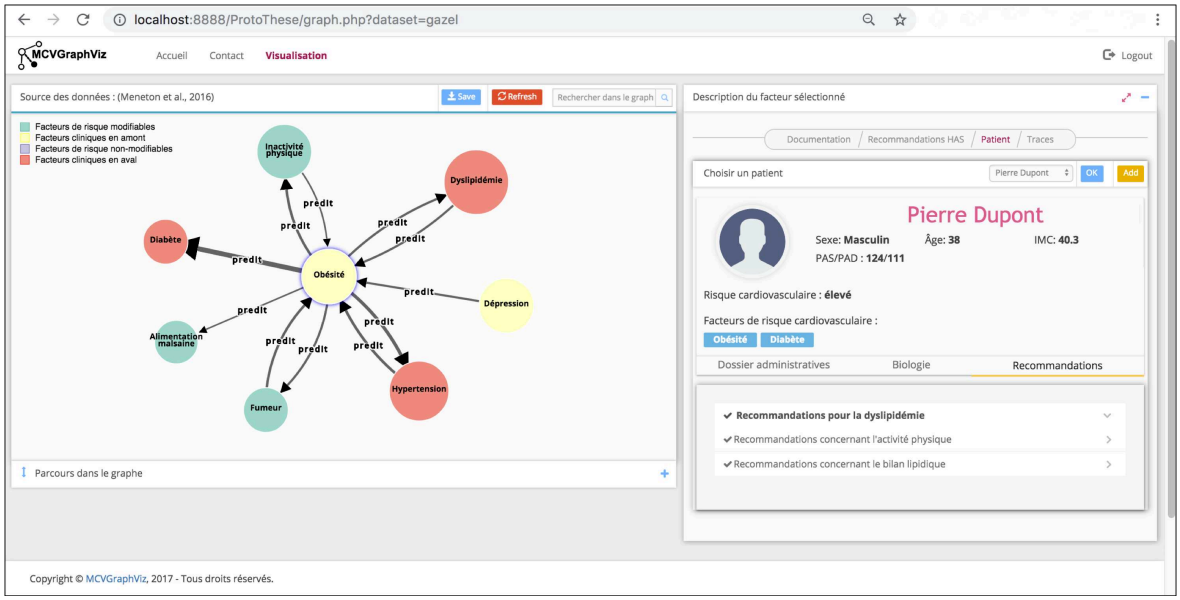


FIGURE 5.5 – Intégration des recommandations dans l'outil MCVGraphViz : accès médecin

chage des informations nécessaires pour aider l'utilisateur et le médecin à prendre une décision, car la décision revient toujours à l'utilisateur et l'inclure dans ce choix est très important pour la réussite de la recommandation. Les utilisateurs et les médecins ont des expertises différentes. D'un côté le médecin qui possède les informations concernant la maladie, le risque, etc. De l'autre, l'utilisateur qui possède les informations concernant son corps, ses conditions de vie, etc. C'est seulement en collaborant et en prenant les décisions ensemble que les interventions efficaces fondées sur les preuves peuvent être atteintes. Ce processus de prise de décision est construit sur l'échange d'informations à propos des options possibles et d'identifier les conséquences potentielles de chacune d'entre elles. L'approche sera d'**informer** l'utilisateur sur les **options possibles** et de prendre une **décision en accord** avec les possibilités du patient.

5.5 Conclusion

Dans ce chapitre, nous avons présenté les systèmes de recommandation en général, puis spécifiquement dans le domaine de la santé. Nous avons comparé la littérature récente sur les systèmes de recommandation en santé, ce qui nous a conduit à identifier un ensemble de critères (liste non exhaustive) essentiels pour concevoir un système de recommandation efficace. L'approche choisie s'appuie sur les recommandations du PNNS concernant l'activité physique et l'alimentation, et les recommandations de la HAS pour une meilleure prise en charge du patient. Comme nous l'avons montré dans ce chapitre, l'alimentation est impliquée dans différents facteurs de risque cardiovasculaire. Ainsi, il est essentiel d'avoir une qualification nutritionnelle précise des recettes consommées, ce qui permet : (i) aux médecins d'avoir un suivi à court et long terme et de mesurer l'impact de la stratégie mise en place avec le patient ; (ii) aux utilisateurs sains de voir l'évolution du profil en suivant des recommandations adaptées. Dans le chapitre suivant, nous décrivons l'approche mise en place pour obtenir une qualification nutritionnelle des recettes avec une traduction visuelle des valeurs nutritionnelles des recettes selon l'échelle du PNNS.

Chapitre 6

Qualification nutritionnelle des recettes de cuisine

Sommaire

6.1 Introduction **128**
6.2 Systèmes de profilage nutritionnel **129**
6.2.1 Définition 129
6.2.2 Eléments de base des systèmes de profilage nutritionnel 130
6.3 Qualification nutritionnelle des recettes de cuisine **131**
6.3.1 Définition 131
6.3.2 Connaissances du domaine 133
6.3.3 Tables et base de données sur la composition des aliments ... 134
6.3.4 Calcul des valeurs nutritionnelles des recettes 136
6.3.5 Traduction visuelle des valeurs nutritionnelles 137
6.4 Approche proposée **138**
6.4.1 Description des ressources utilisées 138
6.4.2 Description des étapes de la méthodologie 143
6.4.3 Enrichissement lexical des produits composant les ingrédients 144
6.4.4 Génération du fichier de calcul 147
6.4.5 Calcul et attribution du score 148
6.4.6 Traduction du score sur l'échelle graphique 150
6.5 Evaluation des résultats **150**
6.6 Conclusion **153**

L'objet de ce chapitre est la présentation du problème de la qualification nutritionnelle des recettes de cuisine qui consiste à automatiser l'attribution d'un score nutritionnel à une recette et son implémentation. L'objectif du travail présenté dans ce chapitre est de construire un module de qualification nutritionnelle des recettes qui s'intègre dans l'écosystème de prévention du risque cardiovasculaire, pour fournir aux utilisateurs, des contenus et des outils pour mieux les renseigner et les guider dans l'adoption d'un mode de vie sain. L'approche adoptée se décompose en quatre étapes : enrichissement lexical en annotant les termes désignant les produits composant les ingrédients ; génération du fichier de calcul à partir d'un patron lexical et interrogation de la table de composition ; calcul et attribution du score ; traduction du score sur une échelle graphique. Cette méthodologie est fondée sur l'appariement entre ressources textuelles (corpus de recettes) et des données structurées (table de composition Nutrinet). La ressource termino-ontologique MIAM est utilisée pour améliorer la qualité

de l'appariement et permettre la caractérisation nutritionnelle des recettes.

Textes de références :

R. Azzi, S. Despres et J. Nobécourt. (2016). Approche sémantique pour automatiser le calcul des valeurs nutritionnelles d'une recette de cuisine.Atelier IA & Santé, IC20016,Jun 2016, Montpellier France.

R. Azzi, S. Despres, G. Guezennec et J. Nobécourt. (2016). Utilisation de ressources sémantiques pour l'automatisation du calcul d'un score nutritionnel.RJCIA2016, Jun 2016, Clermond-Ferrand, France.

6.1 Introduction

Le cadre méthodologique dans lequel nous nous situons est celui présenté dans le chapitre 5. Pour pouvoir mettre en place un système de recommandation alimentaire pour la prévention des maladies cardiovasculaires ou d'autres maladies chroniques liées à l'alimentation (car l'alimentation est liée à plusieurs maladies chroniques). Il est nécessaire d'avoir une qualification nutritionnelle des recettes préparées par l'utilisateur, pour pouvoir évaluer ces apports et mettre en place des stratégies de prévention efficaces. Souvent, la qualification nutritionnelle est réalisée par des nutritionnistes, ce qui implique un temps de travail conséquent vu la quantité énorme de recettes. L'objectif est d'automatiser cette démarche experte et de fournir un service qui pourra être intégré dans le système de recommandation.

Au cours des 30 dernières années, de nombreux travaux scientifiques fondamentaux, cliniques et épidémiologiques ont montré que l'alimentation et les comportements alimentaires constituent des déterminants majeurs pour de nombreuses maladies chroniques non transmissibles (Hercberg, 2013). En effet, jusqu'à 90% des maladies coronariennes, 90% du diabète de type et 30% des cancers pourraient être évités par un mode de vie adéquat, dont la nutrition est reconnue comme un problème critique. Au cours de la même période, l'obésité a considérablement augmenté dans la plupart des pays, y compris chez les enfants, ce qui s'est accompagné d'une augmentation marquée du diabète de type et du syndrome métabolique.

La réponse des autorités a été de mettre en place de nombreux modèles de formation des nutriments pour l'éducation et l'alimentation, la réglementation des allégations de santé et l'évaluation de la qualité nutritionnelle des produits alimentaires (Quinio *et al.*, 2007).

Au niveau international, la Commission de l'Organisation mondiale de la santé (OMS) a proposé certaines séries d'actions de recommandation afin de lutter efficacement contre l'obésité infantile par exemple. L'une de ces actions consiste à développer des modèles de profils nutritionnels basés sur des critères nutritionnels objectifs, transparents et reproductibles, afin de déterminer si un produit alimentaire est éligible ou non éligible pour être commercialisé auprès des enfants.

Un certain nombre de pays, dont la France, le Chili, le Danemark, l'Irlande, le Mexique, la Nouvelle-Zélande, la Corée du Sud, Singapour et le Royaume-Uni, ont déjà pris des mesures en ce sens et utilisent des modèles de profilage nutritionnel (NP) pour réglementer différentes formes (U. Lehmann *et al.*, 2017). Les autorités françaises ont inscrit dans la loi de modernisation du système de santé du 26 janvier 2016 la possibilité de recommander un système d'étiquetage nutritionnel. L'objectif de la mise en place d'un système d'étiquetage nutritionnel simplifié pour les produits alimentaires est d'aider les consommateurs à faire des choix alimentaires plus sains. Voilà pourquoi, le logo Nutri-Score a été conçu par Santé publique France, à la demande de la direction générale de la santé, en s'appuyant sur les travaux de l'équipe du professeur Serge

Hercberg (Hercberg, 2013) qui sont fondés sur le Système de profil nutritionnel de la Food Standards Agency (FSA).

Pour la santé, ces stratégies publiques sont essentielles pour guider les choix des consommateurs et produire un impact substantiel sur la prévention des maladies cardiovasculaires grâce à la nutrition. Cela est dû au rôle important du mode de vie, y compris la nutrition dans l'étiologie des maladies cardiovasculaires. En France, l'étude de l'association potentielle entre le score alimentaire du système de profilage des nutriments (FSA-NPS DI) proposé par la Food Standards Agency et le risque de maladie cardiovasculaire a été déjà abordée dans les travaux de (Adriouch, Julia, Kesse-Guyot, Méjean *et al.*, 2016 ;Adriouch, Julia, Kesse-Guyot, Ducrot *et al.*, 2017). Le résultat de ces travaux montre qu'un choix alimentaire plus médiocre pourrait être associé à une augmentation significative du risque cardiovasculaire, en particulier chez les personnes à risque (fumeurs et personnes physiquement inactives). A partir de ce constat, l'exploitation de ce score permetterait la mise en place de stratégies de prévention en santé publique.

6.2 Systèmes de profilage nutritionnel

6.2.1 Définition

Le profilage des nutriments a aujourd'hui un nouveau sens. Popularisé par la Food Standards Agency, le profilage des nutriments fait désormais référence au jugement qui est porté sur la « santé » relative d'un aliment, généralement révélée par sa composition nutritionnelle (Bussell, 2005).

Le profil nutritionnel (PN) est décrit comme la science du classement ou de la classification des aliments en fonction de leur composition nutritionnelle dans le but de prévenir les maladies et de promouvoir la santé. Les modèles de profil nutritionnel peuvent être utilisés pour diverses applications, y compris la réglementation des allégations nutritionnelles et de santé (WHO, 2011), l'étiquetage des emballages (Roodenburg *et al.*, 2011), etc. Un exemple d'un tel système serait l'étiquetage nutritionnel placé à l'avant des emballages alimentaires, suggérant aux consommateurs qu'ils pourraient manger ces aliments en quantité abondante (vert), avec modération (orange) et de façon mesurée (rouge).

L'OMS définit le terme « profil nutritionnel » comme « la science consistant à classer les aliments en fonction de leur composition nutritionnelle pour des raisons relatives à la prévention des maladies liées à la nutrition et à la promotion de la santé ».

Les profils nutritionnels des aliments sont produits par un ensemble de règles et de formules appliquées aux données sur la composition des aliments (schéma ou modèle de profil nutritionnel). Les PNs seront utilisés dans le cadre du nouveau règlement européen sur les allégations nutritionnelles et de santé afin de différencier les aliments autorisés à porter des allégations relatives à la santé et à la nutrition et les aliments non autorisés (Quinio *et al.*, 2007). Certains schémas ou modèles existent déjà pour différentes applications :

- aider les consommateurs à faire des choix alimentaires raisonnés ;
- identifier les produits éligibles aux allégations santé ;
- prendre des décisions dans les programmes de publicité ou marketing destinés aux enfants ;
- évaluer la qualité nutritionnelle des produits alimentaires.

6.2.2 Eléments de base des systèmes de profilage nutritionnel

Le profilage nutritionnel permet d'établir le PN d'un aliment en le décrivant à partir des nutriments entrant dans sa composition. Des allégations nutritionnelles [1] reposant sur des seuils préétablis sont formulées pour le qualifier. Par exemple, le PN d'une baguette est : 55g pour 100g de glucides complexes, 0,8g pour 100g en lipides. Une allégation nutritionnelle lui correspondant sera : riche en glucides complexes, teneur faible en lipides. De nombreux systèmes de profilage nutritionnel ont été développés dans le monde entier. Ils varient en grande partie dans le format, le contenu et le but pour lequel ils ont été développés (WHO, 2011).

Plusieurs modèles de profil nutritionnel sont actuellement disponibles. Alors que la plupart sont dans le domaine public (Hawley *et al.*, 2012 ; Julia *et al.*, 2015 ; Darmon *et al.*, 2017) certains sont propriétaires (Katz *et al.*, 2009). Ces modèles ont tendance à varier considérablement dans leur sélection de nutriments, la base utilisée pour le calcul (par 100g/kcal ou par portion), et le choix des seuils ou de l'algorithme de notation utilisé (Rayner *et al.*, 2013). Dans les modèles « transversaux », le même algorithme est appliqué à tous les aliments et toutes les boissons, alors que les modèles « spécifiques à une catégorie » utilisent des algorithmes différents pour différentes catégories d'aliments (Nijman *et al.*, 2006). Indépendamment des paramètres utilisés par chaque modèle, les modèles de profil nutritionnel doivent être spécifiquement conçus pour le but pour lequel ils seront utilisés. Il a été démontré que les systèmes d'étiquetage sur le devant du produit avaient un effet sur la reformulation des produits (Vyth *et al.*, 2010).

Deux types de systèmes sont à distinguer : systèmes à définition de seuil (exemple, système feux tricolores) et systèmes avec score (exemple, échelle nutritionnelle du PNNS).

— **seuil** : dans le cas des systèmes à base de seuil, le principe consiste à calculer des seuils pour chacun des nutriments (par exemple, matières grasses, acides gras saturés, etc.) afin de les classer selon une échelle d'intensité en trois couleurs : rouge (élevé), orange (moyen) et vert (faible). La classification de chaque nutriment dans l'une des trois couleurs est fonction de sa teneur dans l'aliment pour 100g ou 100ml ou par portion si la portion est supérieure à 100g ou 150ml. Parmi les systèmes à base de seuil, nous pouvons citer le système d'information « Feux tricolores (Traffic Lights) » mis en place au Royaume-Uni en 2013 ;

— **score** : consiste à calculer un score nutritionnel unique représentant la qualité globale d'un aliment. Le principe consiste à calculer la teneur pour 100g de chaque élément composant l'aliment et attribuer des points selon une échelle à chaque élément en fonction des quantités présentes. Dans certains cas, le score obtenu peut être minimisé par la soustraction d'éléments nutritionnels considérés comme positifs (par exemple, les fruits et légumes). Parmi ces systèmes, nous pouvons citer l'« Échelle nutritionnelle » proposée par Serge Hercberg (Hercberg, 2013) en France en 2013 qui tient son origine du système britannique de profilage des nutriments mis au point par la Food Standards Agency (FSA). Le calcul du score de la FSA comporte trois étapes : (1) attribution d'une note à chaque nutriment en fonction de sa teneur dans l'aliment ; (2) calcul d'un score par une addition des notes obtenues pour les différents nutriments. Il est également possible de calculer deux scores indépendants, l'un à partir des teneurs en nutriments à encourager (fibres, protéines, etc.), l'autre à partir des nutriments à limiter (graisses saturées,

1. http://eur-lex.europa.eu/legal-content/fr/ALL/?uri=CELEX:32006R1924

sel, etc.), ce qui évite la compensation entre les deux ; (3) comparaison du ou des scores globaux à des valeurs repères pour déterminer si l'aliment a accès aux allégations.

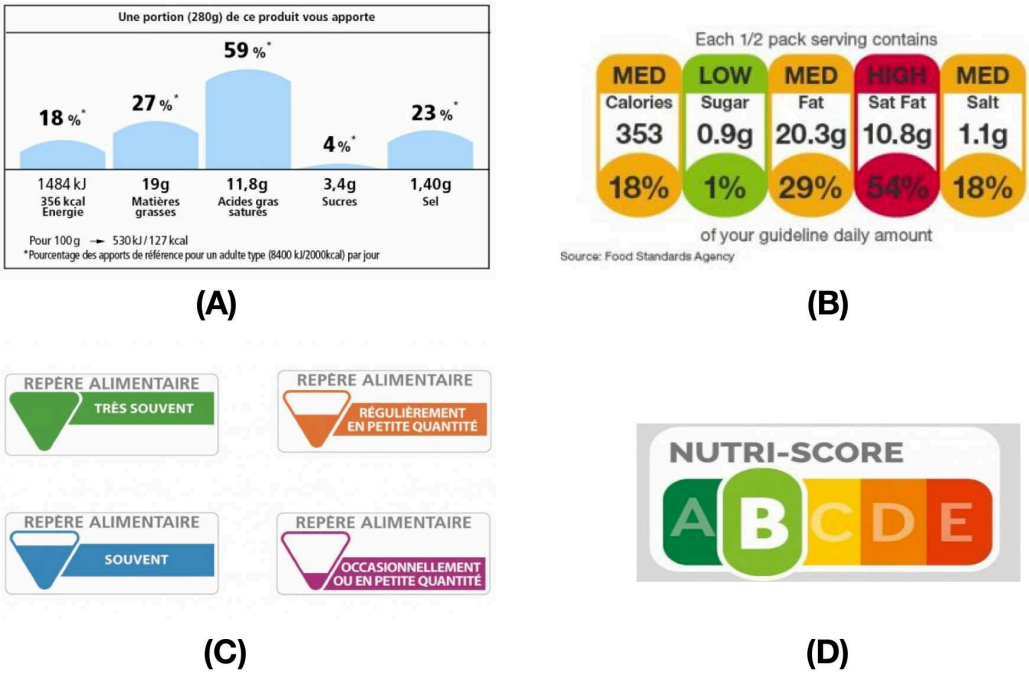


FIGURE 6.1 – Exemple de système de représentation graphique : (A) le système Feux tricolores, créé en 2013 au Royaume-Uni ; (B) le système GDA (Guideline Daily Amounts), créé en 2006 et reconnu en 2011 en Europe ; (C) le système SENS, créé en 2015 en France ; (D) le Nutri-Score créé en 2013 en France.

Les systèmes de représentation graphique des informations nutritionnelles sont, parmi les systèmes fondés sur l'idée qu'il est nécessaire de fournir davantage d'informations aux consommateurs, et de mieux les orienter, pour que se déploient dans la population des comportements alimentaires plus favorables à la santé. Sur ce principe, de nombreux pays ont développé leurs propres systèmes. La figure 6.1) présente quelques-uns de ses systèmes.

6.3 Qualification nutritionnelle des recettes de cuisine

6.3.1 Définition

Le terme recette a plusieurs significations selon le contexte. Il peut être défini d'une manière générale comme une méthode pour aboutir à une finalité souhaitée. Lorsqu'il est utilisé dans le contexte de la cuisine, il est généralement considéré comme un ensemble d'instructions associées à une préparation culinaire. En tant que tel, il pourrait être considéré comme un objet avec des propriétés telles que la liste des ingrédients, le temps nécessaire, etc. Alternativement, il pourrait être considéré comme un processus, prenant en compte certaines entrées et qui se déroulent selon une série d'étapes à exécuter conduisant à la réalisation de cette recette. D'autres informations sont associées aux recettes pour mieux les décrire. Par exemple, le temps nécessaire pour exécuter la recette et les ustensiles nécessaires à sa réalisation. Une recette de cuisine, en dehors des instructions et des ingrédients, contient parfois des informations permettant de la

classer selon différents types. Une recette peut être qualifiée de végétarienne ou appartenant à un groupe particulier telles les recettes d'été, les recettes d'hiver, etc. Des informations spécifiques (telles que la valeur de l'énergie, la valeur de la matière grasse, etc.) peuvent également être nécessaires pour utiliser la recette à des fins spécifiques (par exemple la prévention de certaines maladies chroniques).

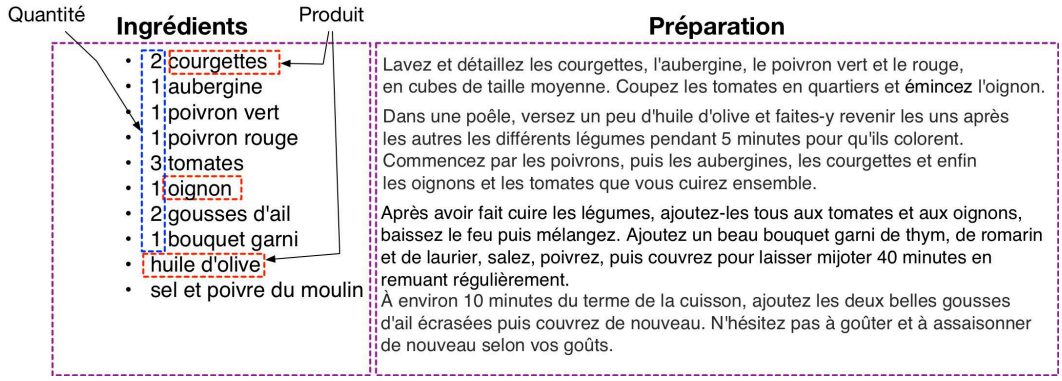


FIGURE 6.2 – Structure d'une recette de cuisine

La figure 6.2 illustre une recette de cuisine, nous pouvons clairement déterminer à quoi correspond une recette de cuisine. Une recette idéale décrit chaque ingrédient dont la description comporte la quantité du produit, l'unité et son nom. Les recettes disponibles en ligne peuvent avoir une quantité définie ou non, une unité peut être définie ou non, mais le nom de l'ingrédient sera toujours mentionné.

Actuellement, de nombreuses applications relatives à la cuisine numérique font leur apparition dans la vie quotidienne. En effet, le développement de l'internet et des nouvelles technologies contribue à l'émergence d'outils permettant de partager avec des amis des coups de cœur culinaires, de créer un livre de recettes personnelles en ligne (DESPRES, 2014). Les métadonnées telles que l'auteur, le titre, liste d'ingrédients, la méthode de préparation, le temps de préparation fournissent des informations pertinentes sur la recette. Le marquage de ces informations peut offrir de nombreux avantages. Si un navigateur web interprète qu'une page web particulière contient une recette avec des caractéristiques spécifiques, il peut permettre des interactions plus riches. Par exemple, des recherches spécifiques d'ingrédients ou d'auteurs peuvent être effectuées *via* des services de recherche généraux tels que Google et Wikipedia. Afin de permettre et d'encourager le partage, la distribution des recettes, les normes microformats [2] sont apparues. Les microformats ont fait leur apparition sur la scène il y a quelques années et deviennent rapidement un outil essentiel pour tous les concepteurs et développeurs Web professionnels. L'objectif est de pouvoir intégrer toutes informations telles que les contacts Web, les articles balisés et les informations géographiques de manière transparente dans des applications Web et de bureau, sans avoir à ajouter quoi que ce soit sur les sites web, à l'exception d'un balisage HTML spécialisé (ALLSOPP, 2007). Dans le cas des recettes, hRecipe [3] est un format simple, ouvert et distribué, adapté pour intégrer des informations sur les recettes de cuisson au format (X)HTML, RSS et XML, etc. La figure 6.3 montre un exemple de recette (ratatouille) décrite à l'aide du microformat hRecipe.

2. `http://microformats.org/wiki/Main_Page`
3. `http://microformats.org/wiki/hrecipe`

```
<div class="hrecipe">
    <h1 class="fn">Ratatouille</h1>
    <p class="summary">
        Très bon chaud mais peut aussi se servir froid.<br />
    </p>
    <p>
        Contributed by <span class="author">Claire Dupont</span> and the
        <span class="author vcard"><a class="url fn" href="http://example.com">Cooky Gang</a></
span>.
    </p>
     <p>Publiée le <span class="published"><span class="value-title" title="2018-10-15"> </span>15.
Oct 2018</span></p>
    <img src="/img/ratatouille.png" class="photo" width="100" height="100" alt= "Ratatouille"/>
    <h2>Ingrédients</h2>
    <ul>
        <li class="ingredient">
            <span class="value">350</span>
            <span class="type">gramme</span> d'aubergine.
        </li>
        <li class="ingredient">
            <span class="value">350</span>
            <span class="type">gramme</span> de courgette.
        </li>
        <li class="ingredient">
            <span class="value">350</span>
            <span class="type">gramme</span> de poivron de couleur rouge et vert.
        </li>
        <li class="ingredient">
            <span class="value">350</span>
            <span class="type">gramme</span> d'oignon.
        </li>
        <li class="ingredient">
            <span class="value">500</span>
            <span class="type">gramme</span> de tomate bien mûres.
        </li>
          …
     </ul>
    <h2>Instructions</h2>
    <ul class="instructions">
        <li>Coupez les tomates pelées en quartier,</li>
        <li>les aubergines et les courgettes en rondelles.</li>
        <li>Emincez les poivrons en lamelles</li>
        <li>et l'oignon en rouelles.</li>
        <li>Chauffez 2 cuillères à soupe d'huile dans une poêle</li>
        <li>et faites-y fondre les oignons et les poivrons.</li>
        <li>Lorsqu'ils sont tendres, ajoutez les tomates, l'ail haché, le thym et le laurier.</li>
        …
    </ul>
    <h2>Plus de détails</h2>
    <p>Nombre de personnes <span class="yield">4</span>.</p>
    <p>Le temps de préparation est d'environ
        <span class="duration"><span class="value-title" title="PT1H20M"> </span>80 minutes</span>
    </p>
    <p>Cette recette est <a href="http://www.example.com/tags/difficulty/easy" rel="tag">
       Facile</a> et <a href="http://www.example.com/tags/tastyness/delicious" rel="tag">
       Délicieuse</a>.</p>
    <p>
        <span class="nutrition">
        Les valeurs nutritionnelles de cette recette :
        <span class="value">54.5</span>
        <span class="type">Kcal</span>Energie</span>,
    </p>
</div>
```

FIGURE 6.3 – La recette de cuisine ratatouille décrite à l'aide du microformat hRecipe

6.3.2 Connaissances du domaine

Il existe de nombreuses sources d'informations sur les aliments et les recettes construites selon des critères qui ne sont pas forcément pertinents pour notre travail. Par conséquent, nous avons étudié les sources de connaissances existantes sur le domaine de la cuisine et a vérifié leur adéquation. Nous pouvons distinguer :

1. les bases de données relatives à la composition des aliments sont généralement publiées avec des formats et des niveaux de documentation différents. Ainsi, il est

souvent difficile de comparer les valeurs ou d'importer des données en utilisant une procédure standard. Par exemple, l'USDA [4] est une base de données de référence créée par le Département de l'agriculture des États-Unis pour constituer la principale source de données sur la composition des aliments aux États-Unis ;

2. les tables de composition des aliments qui fournit les teneurs en différents constituants (glucides, sucres, protéines, lipides, acides gras, sel, vitamines, minéraux, etc.) des aliments, par exemple, la table française Ciqual [5] ;
3. Thesaurus et ontologies, par exemple AGROVOC [6] qui est un thesaurus multilingue établi par l'Organisation des Nations Unies pour l'alimentation et l'agriculture (FAO) qui compte environ 35 000 concepts et types de relations (terme préféré, terme connexe et terme plus large). Le thésaurus multilingue LanguaL [7] utilise la classification à facettes. Chaque aliment est décrit par un ensemble de termes standard et contrôlés. Les facettes sont choisies parmi des facettes qui caractérisent la qualité nutritionnelle ou hygiénique d'un aliment, par exemple l'origine biologique, les méthodes de cuisson et de conservation et les traitements technologiques. (Sam *et al.*, 2014) proposent un modèle de conception d'ontologie pour les recettes de cuisine. L'objectif de cette modélisation était de rapprocher l'hétérogénéité des choix de représentation en développant un modèle de conception d'ontologies de contenu suffisamment général pour permettre l'intégration d'informations provenant de différents sites Web. (Dmitriev, 2018) proposent un cadre pour la représentation formelle de recettes basées sur GSL (Globbie Semantic Language) et stockées dans Knowdy [8], une base de données de graphes open source. L'objectif est de préserver le contexte culturel unique des différentes zones géographiques. (Ribeiro *et al.*, 2006) proposent une ontologie pour le domaine de la cuisine, à intégrer dans un système. L'ontologie comprend quatre modules principaux couvrant les concepts clés du domaine de la cuisine actions, aliments, recettes et ustensiles, et trois modules auxiliaires unités et mesures, équivalences et types de plaques. (Despres, 2014) propose une ontologie modulaire de la cuisine numérique construite à partir des connaissances acquises auprès des experts du domaine (nous reviendrons sur cette ontologie dans les parties suivantes).

6.3.3 Tables et base de données sur la composition des aliments

De nombreux pays ont aujourd'hui leurs propres bases de données nationales sur la composition des aliments, dont la mise à jour permanente permet d'inclure un grand nombre d'aliments, reflétant le mieux possible les habitudes alimentaires de la population et le nombre croissant d'aliments sur le marché. Par conséquent, une attention particulière devrait être portée à l'étude non seulement des aliments ou des composants alimentaires individuels, mais aussi aux caractéristiques nutritionnelles des plats, des repas et des régimes alimentaires, tels qu'ils sont réellement consommés (Marconi *et al.*, 2018).

L'USDA est la base de données sur la composition des aliments aux États-Unis. Elle contient des informations nutritionnelles sur les produits alimentaires, les groupes

4. `https://ndb.nal.usda.gov/ndb/`
5. `https://ciqual.anses.fr/`
6. `http://aims.fao.org/fr/agrovoc`
7. `http://www.langual.org/`
8. est un projet de logiciel open source (`https://github.com/globbie/knowdy`), qui vise à développer une base de données de graphes ultrarapide permettant une manipulation directe et efficace avec des graphes conceptuels, en ignorant toute représentation intermédiaire telle que les tables SQL.

alimentaires et des produits de grande fabrication. Elle contient 247 326 aliments, dans sa version 3.3 qui date de mars 2018.

Le Fichier canadien sur les éléments nutritifs (FCEN [9]) est la base de données de référence sur la composition des aliments qui renseigne la quantité des éléments nutritifs dans les aliments couramment consommés au Canada. Le FCEN est utilisé par tout un spectre d'usagers externes incluant les milieux hospitaliers, les universités, l'industrie alimentaire et le grand public. Dans sa version de 2015, elle contient 5692 aliments.

La « Table de composition nutritionnelle suisse [10] » rassemble les données sur les teneurs en nutriments des aliments les plus divers. Elle est administrée par l'Office fédéral de la sécurité alimentaire et des affaires vétérinaires OSAV. Elle est spécialement fondée sur les habitudes alimentaires de la Suisse et contient des informations nutritionnelles sur des denrées alimentaires suisses.

Actuellement, en France, il existe deux principales tables de composition, la table de référence CIQUAL [11] et la table, NUTRINET [12]. Ces tables sont respectivement produites et maintenues par l'Anses et l'EREN.

La table de composition Ciqual renseigne sur la composition nutritionnelle moyenne des aliments les plus consommés en France. Chaque aliment est décrit par 61 constituants, selon les données disponibles : protéines, lipides, glucides, acides gras, etc. Les teneurs sont toujours fournies pour 100g de portion comestible, c'est-à-dire sans les os pour la viande, sans le trognon pour la pomme, etc. Le Ciqual a été créé en 1985, à l'initiative conjointe des pouvoirs publics et des industries agroalimentaires. Dans sa version de 2017, elle contient 2807 aliments.

La table de composition NUTRINET fournit un ensemble complet de données sur les nutriments spécifiques à l'étude Nutrinet-Santé réalisée pour mieux comprendre les relations entre la nutrition et la santé en France. Elle fournit la composition en 43 nutriments pour approximativement 2122 aliments dans sa première version de 2009. Pour chaque nutriment, les données sont exprimées pour 100g d'aliment tel qu'il est consommé en France. Cette table est construite à partir de différentes tables de composition citées dans la littérature et chaque aliment est validé par les diététiciens de l'EREN. A titre d'exemple, 450 aliments proviennent de la table Ciqual. Contrairement à d'autres tables, NUTRINET contient des informations essentielles pour le calcul des valeurs nutritionnelles des recettes. Par exemple, le poids des aliments après cuisine, le pourcentage des parties comestible, les lipides d'origine animale, les lipides d'origine végétale, etc.

En Europe EuroFIR [13] a également travaillé à la création de normes pour la compilation de bases de données, y compris la mise en forme standard des échanges au sein des pays européens afin de permettre la comparaison et l'harmonisation des bases de données des différents pays (Finglas *et al.*, 2014 ;Ireland et Møller, 2013). Une étude menée par (Ireland et Møller, 2013) a examiné la capacité à normaliser les tables de composition (FCEN, USDA, etc.) par rapport au projet de critère de format de base de données développé par EuroFIR. L'étude a identifié que ces bases de données différaient par la méthodologie utilisée pour définir et décrire les aliments, ainsi que par la manière dont les valeurs nutritionnelles étaient documentées.

9. https://aliments-nutrition.canada.ca/cnf-fce/index-fra.jsp
10. http://www.valeursnutritives.ch
11. https://ciqual.anses.fr/
12. https://www.etude-nutrinet-sante.fr/
13. http://www.eurofir.org/

6.3.4 Calcul des valeurs nutritionnelles des recettes

L'estimation de l'apport en nutriments d'une recette nécessite des données fiables sur la composition des aliments. Ces données constituent également les principes fondamentaux des recommandations alimentaires basées sur l'alimentation pour une alimentation saine, contenant les informations nécessaires sur les sources de nourriture pour différents nutriments (Elmadfa et Meyer, 2010).

Selon (Charrondiere *et al.*, 2011), pour calculer la teneur en éléments nutritifs pour 100g de portion comestible d'aliments préparés (sous forme d'une recette), les données suivantes sont requises : (i) informations de base sur les recettes (par exemple, méthode de cuisson, quantité d'ingrédients, absorption des graisses, rendement pondéral dû à la préparation) ; (ii) composition nutritionnelle des ingrédients par 100g de portion comestible ; (iii) facteurs de rétention des éléments nutritifs après la préparation des aliments (plat).

Cependant, certaines informations relatives au rendement pondéral après la préparation du plat ne sont généralement pas incluses dans les livres de recettes ou sur les applications relatives à la cuisine numérique disponibles sur internet. Par conséquent, le rendement en poids doit être mesuré en pesant la quantité totale des ingrédients et la quantité d'aliments préparés (plat).

6.3.4.1 Méthode de calcul des valeurs nutritionnelles

La méthode standard pour calculer les valeurs nutritionnelles comme les calories pour une recette consiste à analyser chimiquement le résultat final de la recette une fois cuite (Picó, 2012). Cependant, il n'est pas facile d'effectuer une analyse chimique des plats, car cela implique des coûts plus élevés (en termes d'argent et de temps) et cette approche peut nécessiter plusieurs évaluations pour déterminer la composition de chaque élément nutritif. Compte tenu du nombre de recettes disponibles en ligne, cette analyse chimique ne semble pas être une solution pratique pour calculer les valeurs nutritionnelles des recettes.

De nombreuses autres approches ont été proposées, ces méthodes se différencient en fonction de l'utilisation ou non des facteurs de rendement (changement du poids des aliments ou des recettes dû à la cuisson) et de rétention des nutriments (vitamines et minéraux après cuisson, stockage, préparation, réchauffement, etc.) et le niveau auquel intervient ce calcul. Les facteurs sont applicables au niveau de l'ingrédient (c'est-à-dire aux teneurs en nutriments de chaque ingrédient) ou au niveau de la recette (c'est-à-dire après avoir additionné les teneurs en nutriments de tous les ingrédients). Nous pouvons distinguer :

— Méthode « d'addition des valeurs nutritionnelles des ingrédients crus » : les teneurs en nutriments des ingrédients crus sont additionnées sans appliquer les facteurs de rendement et de rétention. Cette méthode n'est pas recommandée à l'exception des plats non cuits (par exemple, salade de fruits) ;

— Méthode « ingrédient » : les facteurs de rendement et de rétention des nutriments sont appliqués à chaque ingrédient, puis les poids des ingrédients cuits sont ajoutés. L'inconvénient est qu'il faut connaître le facteur de rendement de chaque ingrédient ;

— Méthode « recette totale » : les poids des ingrédients crus sont additionnés, le calcul du facteur de rendement est calculé à partir du poids de la recette crue et cuite, l'application des facteurs de rendement et de rétention des nutriments se

fait au niveau de la recette globale. Le facteur de rétention se fait selon l'ingrédient principal ;

— Méthode « mixte » : le calcul du facteur de rendement est obtenu à partir du poids de la recette crue et cuite, le facteur de rendement est appliqué au niveau de la recette et les facteurs de rétention des nutriments au niveau des ingrédients. La plupart des pays utilisent cette méthode.

Comme il est pratiquement impossible d'analyser chaque plat et aliment préparé, le calcul de la teneur en éléments nutritifs reste la méthode de choix. Le meilleur accord entre les données analytiques et les données calculées est obtenu lorsque les changements de poids et de nutriments pendant la préparation (cuisson) sont pris en compte. Le calcul de la composition nutritive est basé sur la quantité d'ingrédients pour la préparation de 100g de portion comestible de plat prêt à servir, sur la base des données relatives à la teneur en éléments nutritifs pour 100g de portion comestible d'ingrédients et de facteurs de rétention. Les informations sur les recettes et l'algorithme nécessaire à la gestion des recettes et au calcul de la composition en nutriments des plats sont décrits et illustrés par des exemples (Bognár et Piekarski, 2000).

Parmi les premiers travaux, (Powers et Hoover, 1989) présentent de nombreux algorithmes improvisés en tenant compte des valeurs nutritionnelles perdues lors d'un processus de cuisson qui pourraient faire varier les calories calculées avant la cuisson et après la cuisson. Ces méthodes doivent également avoir une recette dans un certain format pour calculer les calories. Cependant, les expériences faites par les auteurs ont prouvé que combiner les calories pour les ingrédients avant la cuisson fournissent des résultats acceptables à condition que les ingrédients soient sélectionnés avec précision.

Dans le cadre du projet EuroFIR, la question d'une procédure unique a été discutée, comparée et validée (Bognár et Piekarski, 2000 ; Unwin, 2000). Cette procédure de calcul considère : (i) l'application d'un facteur de rendement au niveau de la recette et des facteurs de rétention au niveau de l'ingrédient ; (ii) l'importance de la sélection de facteurs spécifiques de rétention des nutriments est également soulignée. D'ailleurs, (Westenbrink *et al.*, 2016) a également suggéré l'utilisation de facteurs de rétention communs pour les vitamines et les minéraux selon la classification des aliments EuroFIR et les méthodes de cuisson disponibles dans le thesaurus LanguaL.

D'autres approches ont été proposées, y compris les techniques de reconnaissance d'image, pour déterminer les calories provenant des images de la recette après préparation (Kitamura *et al.*, 2010). Ces techniques essaient d'abord de trouver les composants principaux sur la base de la reconnaissance des formes et d'essayer de prédire les calories en fonction des résultats. Cependant, mis à part les résultats indiquant que les gens ordinaires montrent un intérêt pour l'utiliser, la précision de ces méthodes est faible. Et, la recette doit être cuite avant de calculer les recettes.

6.3.5 Traduction visuelle des valeurs nutritionnelles

Afin de faciliter l'information du consommateur et pour l'aider à choisir en toute connaissance de cause, sans préjudice des dispositions des articles 9, 16 et 30 du règlement (UE) n° 1169/2011 du Parlement européen et du Conseil du 25 octobre 2011 concernant l'information des consommateurs sur les denrées alimentaires, la déclaration nutritionnelle peut être accompagnée d'une présentation ou d'une expression au moyen de graphiques ou symboles au sens de l'article 35 du même règlement. Par exemple, avec ce règlement, l'utilisation d'une procédure de calcul pour formuler les étiquettes

nutritionnelles des produits transformés a été acceptée comme alternative possible à l'analyse chimique directe de recettes.

La fiabilité des calculs des valeurs nutritionnelles de recettes dépend de plusieurs facteurs qui sont la présence de données manquantes, l'absence de facteurs de rétention spécifiques, l'utilisation de bases de données nationales sur la composition des aliments (Reinivuo *et al.*, 2009 ;Greenfield et Southgate, 2003).

6.4 Approche proposée

L'objectif de ce travail est la mise en place d'une méthodologie automatisant le calcul nutritionnel d'un score pour caractériser des recettes (a partir du scénario de l'annexe B) en se fondant sur l'utilisation de ressources termino-ontologique et les tables de composition des aliments. Le calcul de la valeur nutritionnelle d'une recette de cuisine prend en compte pour chacun des ingrédients la manière dont il est préparé et la forme sous laquelle il est utilisé. Le score tient compte d'une part de l'énergie, des graisses saturées, des sucres, du sodium [14], et d'autre part de la proportion des fruits, légumes/noix, des fibres et des protéines [15]. Le score calculé est traduit visuellement sur l'échelle de qualité nutritionnelle (Hercberg, 2013) représentée par cinq disques de couleur couplés à cinq lettres. Le disque correspondant à ce score est mis en évidence graphiquement dans l'échelle par une augmentation de sa surface. L'annotation de ce score explicite les calculs correspondant à la note de couleur obtenue.

Techniquement, l'approche adoptée est incrémentale. Ainsi, toutes les informations ajoutées par l'algorithme restent accessibles à travers les différentes étapes de traitement. L'algorithme opérationnalise des traitements automatiques de la langue (TAL) de base (segmentation, parsing, tokenisation) et de plus haut niveau (extraction d'information). Les principaux traitements sont : (i) segmentation en phrases ; (ii) tokenisation ; (iii) parsing de texte ; (iv) interrogation d'une ressource termino-ontologique ; (v) extraction d'informations ; (vi) analyse morphologique (reconnaissance des aliments dans la base de données). Pour réaliser les étapes de la méthodologie, nous avons utiliser les ressources suivantes : (1) corpus de recettes ; (2) tables de composition des aliments ; (3) ressource termino-ontologique.

6.4.1 Description des ressources utilisées

6.4.1.1 Description d'une recette

La recette est une donnée semi-structurée au format XML. Cette représentation des données a été obtenue par des traitements de mise en forme de texte brut de recettes. Les éléments de la recette pris en compte (voir figure 6.4) sont : (1) la liste des ingrédients ; (2) la liste des instructions relatives à la préparation de la recette.

A la lecture d'une recette, nous observons que les ingrédients ne figurent pas toujours explicitement dans la partie préparation ce qui rend difficile leur reconnaissance dans la préparation. En effet dans certains cas les ingrédients sont mentionnés dans la préparation sous forme : (1) d'un verbe (poivrer, au lieu de poivre) ; (2) d'un raccourci (filet de poisson désigné par filet) ; (3) implicite (l'instruction faire revenir les oignons dans la préparation, déduire que c'est dans l'ingrédient huile d'olive) ; (4) d'un mélange (mélanger les ingrédients de la pâte, etc.). En outre, certains ingrédients apparaissant

14. les teneurs élevées sont considérées comme mauvaises pour la santé
15. les teneurs élevées sont considérées comme bonnes pour la santé

```
=ingredients
échalote                3     pièce
poivron                 2     pièce
farine                  5     cas
sel                           PM
poivre                        PM
huile                   2     cas
poulet en morceaux      1     pièce
poulet                  1     pièce
tomate pelé             1,25  boîte
=instructions
Pelez et hachez les échalotes. Epépinez puis émincez les poivrons. Assaisonnez et
farinez les morceaux de poulet.
Dans l'autocuiseur, chauffez l'huile et faites colorer les morceaux de poulet. Ajoutez
l'échalote et les poivrons en remuant, puis les tomates non égouttées. Assaisonnez.
Fermez l'autocuiseur.
Dès que la vapeur s'échappe, baissez le feu et laissez cuire selon le temps indiqué.
Ouvrez l'autocuiseur. Parsemez de persil haché. Servez.
```

FIGURE 6.4 – Structure d'une recette de cuisine

dans la préparation ne figurent pas dans la liste initiale (suggestions d'accompagnement pour le plat). Nous avons effectué une analyse décrite dans la figure 6.5 pour d'identifier les types de structures à prendre en considération. Ce processus a été expérimenté sur un sous-corpus de 150 recettes de cuisine (nous disposons d'un corpus de 30000 recettes) pour évaluer le pourcentage d'apparition des ingrédients dans la préparation (AIP). Les recettes ont été découpées en ingrédients et préparation. Le découpage a produit 150 listes d'ingrédients et 150 listes d'instruction de préparation. Ensuite, nous avons calculé le pourcentage d'AIP à l'aide de la formule :

$$\% \text{ AIP} = \frac{\sum_{i=1}^{n} (\frac{\text{nombre d'ingrédients reconnus}}{\text{nombre total d'ingrédients}})}{\text{nombre total de recettes}} \tag{6.1}$$

Le résultat obtenu concernant le pourcentage d'AIP été de 45%. Une grande partie des ingrédients non reconnus dans la préparation (55%) sont dominés par des phénomènes liés à la substitution des ingrédients. Par exemple, l'ingrédient « poivre » est remplacé par le verbe « poivrer », ou bien l'ingrédient « feuille de laurier » est cité comme « laurier ». Ce pourcentage faible peut s'expliquer par l'origine des recettes qui sont rédigées par des internautes sans aucune contrainte sur le web.

6.4.1.2 Description de la table de composition des aliments

La plupart des nutriments essentiels sont communs à toutes les tables de composition. Les différences se situent au niveau de la couverture non exhaustive des aliments par les tables. Un fait troublant pour l'utilisateur est que, pour un même couple aliment-constituant, les valeurs indiquées dans deux tables peuvent être très différentes (voir tableau 6.1). Ce fait s'explique assez souvent par l'origine des produits consommés d'un pays à l'autre. Par exemple, le kiwi consommé aux États-Unis est très différent (espèce, lieu de production, etc.) de celui consommé en France. Cependant, le fait que les valeurs indiquées soient identiques ne doit pas obligatoirement être interprété comme favorable. Plus les données sont rares et chères et plus les emprunts deviennent fréquents.

« Pour un consommateur français, le mot pain désigne un aliment légalement fabriqué à partir de farine blanche de blé panifiable, de levure, de sel et d'eau, à l'exclusion de tout autre composant. Si l'on en reste à des dictionnaires linguistiques anglais ou allemand, on peut croire que les termes bread et Brot désignent bien le même aliment.

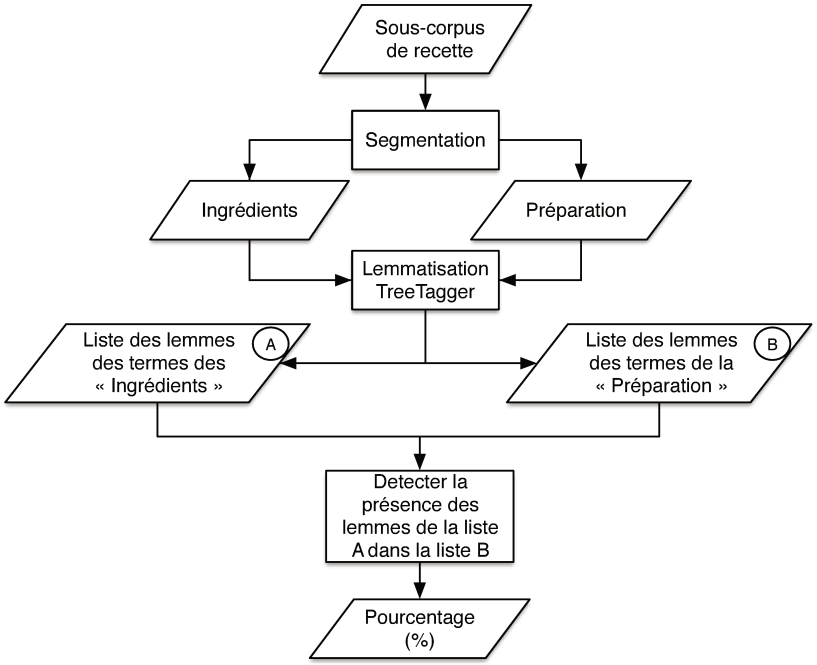


FIGURE 6.5 – Calcul du pourcentage d'apparition des ingrédients dans la préparation de la recette

Aliment	**Table**	Énergie	Protéines	Lipides	Sodium
Kiwi	Nutrinet	49.7	1.14	0.61	1.65
	USDA	60	1.23	0.56	3
Caviar	Nutrinet	253	25	17	1700
	USDA	264	24.6	17.9	1500

TABLE 6.1 – Comparaisons de quatre valeurs nutritionnelles de deux aliments de Nutrinet et USDA

Par contre, si l'on regarde les législations de chaque pays, on s'aperçoit alors que ces dénominations peuvent s'appliquer à des produits de compositions bien différentes de celle du pain français. Il est donc clair que le transfert de données obtenues dans d'autres pays ou à partir de données étrangères exige beaucoup de prudence. C'est pourquoi, dans un système moderne de gestion des données nutritionnelles, à côté du nom de chaque aliment, il est indispensable de faire figurer une description complète, précise et scientifique » (FEINBERG *et al.*, 1992).

Nous avons finalement sélectionné la table Nutrinet (voir figure 6.6). Elle répond le mieux aux critères de qualité définie par (GREENFIELD et SOUTHGATE, 2007) :

1. Les données devraient être représentatives : Les valeurs devraient représenter la meilleure estimation possible de la composition habituelle des aliments, sous les formes les plus couramment disponibles ou consommées.
2. Les données devraient être issues d'analyses de bonne qualité : L'idéal est de disposer de données analytiques originales provenant de sources rigoureusement vérifiées.
3. La couverture des aliments devrait être exhaustive : La base de données devrait inclure tous les aliments représentant une part importante des aliments les plus consommés et, autant que possible, les aliments consommés moins fréquemment.
4. La couverture des nutriments devrait être exhaustive : Les données relatives à

tous les nutriments et autres constituants qui sont ou seraient importants pour la santé humaine devraient être incluses. Les priorités nationales concernant la santé jouent un rôle majeur dans la sélection des nutriments à inclure.

5. Les descriptions des aliments devraient être claires : Pour être facilement identifiés, les aliments doivent être nommés et décrits de façon non ambiguë.

6. Les données devraient être exprimées de manière uniforme et sans ambiguïté : Le mode d'expression des données devrait permettre de les exprimer sans ambiguïté. L'emploi des unités, l'utilisation de facteurs dans les calculs et les arrondissements de valeurs doivent être cohérents.

7. L'origine des données devrait être indiquée au niveau des valeurs nutritionnelles : Il faudrait fournir des informations sur les sources des données, en particulier indiquer s'il s'agit de données analytiques, calculées ou imputées et, s'il y a lieu, sur les procédures de calculs et d'imputations, et les méthodes d'échantillonnage et d'analyse.

8. Les tables et les bases de données devraient être faciles à utiliser : Outre le fait d'utiliser une terminologie claire et une présentation systématique, les bases de données et les tables informatisées devraient être facilement accessibles et rapidement comprises.

9. Le contenu des différentes bases de données devrait pouvoir être compatible : La description des aliments, les modes d'expression et les sources des valeurs doivent être aussi conformes que possible aux normes internationales existantes (par exemple les identificateurs des composants de INFOODS) et proches d'autres bases de données importantes sur la composition des aliments.

10. La bases de données devrait avoir peu de données manquantes : On peut déduire de ce qui précède que toute base ou table de données sur la composition des aliments devrait viser à avoir aussi peu de vides que possible, car les données manquantes pourraient fausser sensiblement les estimations d'apports en nutriments. D'une part, il vaut sans doute mieux inclure des données imputées ou empruntées, toujours clairement identifiées comme telles, plutôt que laisser un vide.

En outre, Nutrinet présente l'avantage d'être en français, de répertorier des produits consommés et achetés sur le marché français. D'ailleurs, elle est issue d'une compilation de l'enquête Nutrinet-Santé menée sue le territoire français et correspond donc aux habitudes alimentaires des consommateurs français. Du point de vue nutritionnel, elle présente les avantages suivants : (i) les teneurs minimales et maximales sont précisées ; (ii) les parties comestibles sont renseignées ; (iii) le taux de sodium après cuisson est renseigné ; (iv) les lipides d'origine animale et végétale sont différenciés ; (v) les taux des lipides et sucres ajoutés sont renseignés.

De plus, nous avons constitué une base de données (BD) relationnelle à partir de la table Nutrinet, dans laquelle chaque aliment est décrit selon les facettes du thésaurus LANGUAL (Azzi, 2014). L'objectif de cette transformation est d'améliorer la qualité des données, en augmentant le niveau de précision des informations accessibles. Nous pouvons alors effectuer des traitements impensables par des moyens manuels. Par exemple, sélectionner tous les aliments dont la valeur en énergie est comprise entre 350 et 380 kcal.

1	libelle	code_aliment	code_marque	Kcal	Prot	ProtV	ProtA	Lip
14	ail frais	660	-1	92,6	7,1	7,1	0	0,5
15	aile de dinde rôtie	88	-1	211,4	29,81	0	29,81	11,701
16	aile de poulet rôtie	3269	-1	282,58	26,86	0	26,86	19,46
17	algue, nori, crue	3836	-1	45	5,81	5,81	0	0,28
18	algue, nori, séchée	3849	-1	247,03	36,22	36,22	0	1,75
19	algue, varech, crue	3850	-1	44,84	1,68	1,68	0	0,56
20	aligot / truffade	1280	-1	197,9769	7,79082	1,2163	6,57452	13,43778
21	amandes grillées	3509	-1	593	23,6	23,6	0	52,9
22	amandes non salées	480	-1	593	23,6	23,6	0	52,9
23	amandes salées	3272	-1	575,5	19	19	0	53,5

FIGURE 6.6 – Table de composition des aliments nutrinet

6.4.1.3 Description de la ressource termino-ontologique

Nous avons choisi d'utiliser l'ontologie (exprimée en OWL 2) modulaire de la cuisine numérique (DESPRES, 2014) qui comporte sept modules : CUISINE, PREPARATION, ALIMENT, UNITE, MATERIEL, NUTRITION et PERSONNE. Seuls les modules ALIMENT et PREPARATION sont utiles à notre travail. Le module ALIMENT compatible LANGUAL comporte les concepts (environ 11035) relatifs aux aliments, leurs modes de conditionnement et leurs traits caractéristiques. Ce module permet de répondre aux interrogations concernant les aliments indépendamment de leur utilisation en cuisine. La figure 6.7 présente une description de ce module. Le module PREPARATION (environ 6164 concepts) est relatif aux préparations de base (par exemple, fond de cuisson, élément de liaison, etc.) utilisées dans une recette. Il décrit les transformations d'aliments *via* les actions culinaires et les types de cuisson.

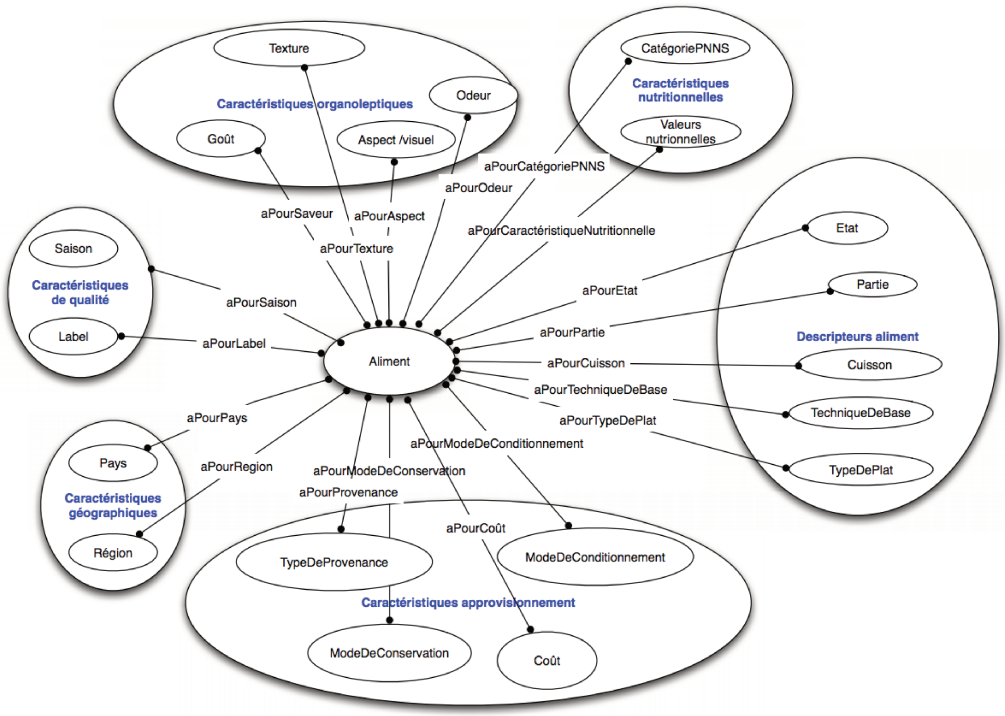


FIGURE 6.7 – Modèle des aliments issu de (DESPRES, 2014)

Notre méthodologie d'automatisation du calcul du score nutritionnel est constituée de quatre étapes (voir figure 6.8) : (1) enrichissement lexical par ajout des termes aux produits composant les ingrédients ; (2) génération du fichier de calcul ; (3) calcul et attribution du score ; (4) traduction du score sur l'échelle graphique.

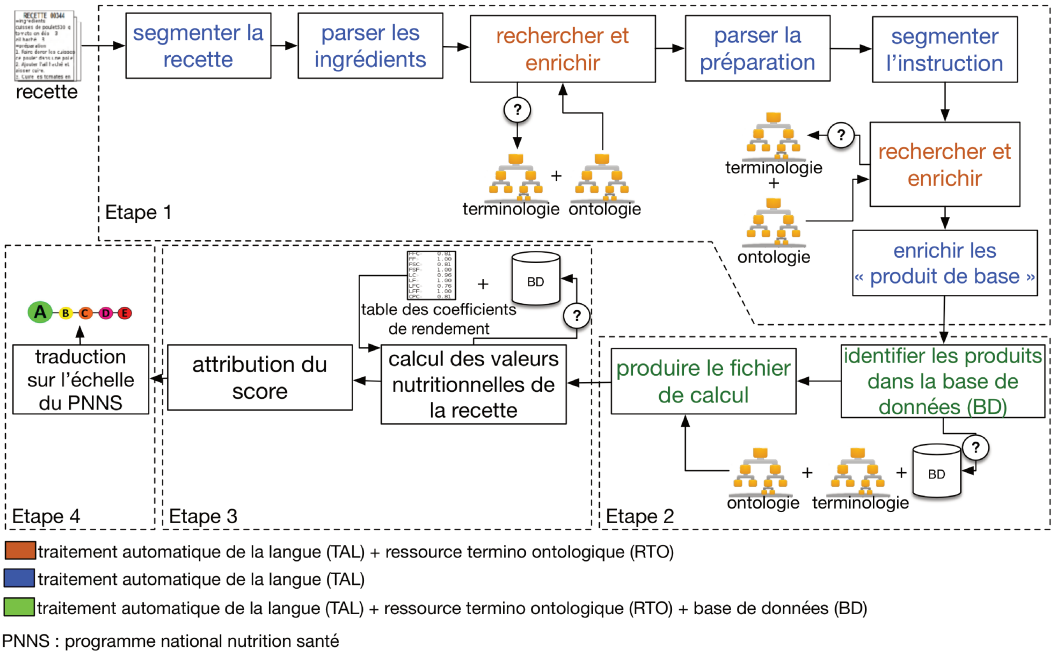


FIGURE 6.8 – Etapes de la méthodologie de calcul des valeurs nutitionnelles des recettes

6.4.2 Description des étapes de la méthodologie

Pour exécuter les étapes de la méthodologie d'une manière optimale, nous avons adopté une méthode de traitement à deux couches symboliques : une couche lexicale et une couche de règles. L'objectif est de gérer les cas précédemment identifiés :

— aller au-delà d'une simple recherche lexicale. Par exemple, dans l'expression « cuisse de poulet », seul « poulet » est un « produit », alors que « cuisse » est une « partie de produit » ;

— l'extraction d'ingrédients « implicites » à partir de certains verbes, par exemple, l'action « fariner » fait intervenir l'ingrédient « farine » ;

— l'expansion des formes. Par exemple, « chantilly [16] » donne « crème chantilly [17] » ;

— l'expansion de termes génériques. Par exemple, le mot « sauce » n'est pas un produit en soi (il peut correspondre à un ingrédient dont les VNs sont renseignés au niveau de la table de composition des aliments ou bien à un aliment composé de plusieurs ingrédients référencés ou non dans la liste d'ingrédients de la recette).

Une étape préliminaire au traitement de la recette permet d'extraire la liste des ingrédients qui ont servi pour réaliser le plat.

Divers outils et langages sont utilisés pour récupérer les données textuelles sémantiquement pertinentes du modèle ontologique. Le langage d'interrogation SPARQL est la méthode la plus courante pour récupérer les données textuelles stockées dans l'ontologie (MAGESH, 2013). La recherche sémantique est différente de la recherche par mot-clé ordinaire. La recherche sémantique (HARISH *et al.*, 2011) consiste en la construction d'un moteur de requête qui reçoit des requêtes dans un langage de requête ontologique (comme SPARQL), les exécute sur la structure ontologique et renvoie des tuples de valeurs satisfaisant les conditions de la question. Cette méthode est une recherche booléenne sur un graphe RDF. Comme il s'agit d'une structure ontologique formelle,

16. désigne un label dans la RTO
17. désigne un altlabel dans la RTO

elle évite l'ambiguïté et la redondance. La signification sémantique est identifiée en effectuant un retour en arrière sur la hiérarchie des classes. Le raisonneur est utilisé pour vérifier la cohérence des données dans une ontologie et effectue un raisonnement sémantique dans l'ontologie (Fadzli et Setchi, 2010).

La partie la plus difficile de l'écriture d'une requête de type SPARQL consiste à composer une clause de filtrage (introduite par le mot-clé *Where*) qui nécessite une compréhension complète des concepts inclus et de leurs relations par diverses propriétés, ce qui a conduit à la mise en place d'autres approches pour faciliter la tâche des utilisateurs lors de l'écriture de ces instructions de requête. Par exemple, *Network Query Language (NQL)* utilise un algorithme de génération d'instruction SPARQL pour générer dynamiquement l'instruction en fonction de l'entrée de l'utilisateur et de l'état de base RDF (He *et al.*, 2016).

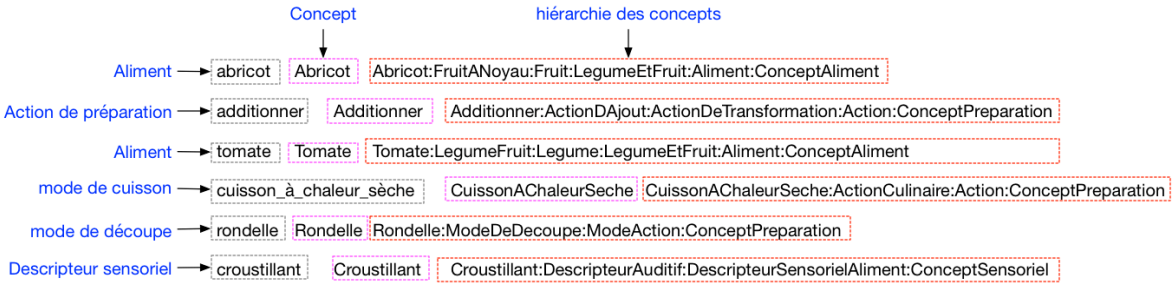


Figure 6.9 – Extrait de la structure du lexique construit à partir de la RTO

Pour la couche de traitement lexical, nous avons construit un lexique (voir figure 6.9) de tous les labels et altlabels présents dans la RTO. Ainsi, nous pouvons effectuer des interrogations sur ce lexique grâce à un algorithme construit pour ce besoin.

6.4.3 Enrichissement lexical des produits composant les ingrédients

La première étape est une procédure d'analyse de la recette aboutissant à l'enrichissement sémantique des unités lexicales (voir annexe B). Elle identifie les unités lexicales, explicite leurs natures (aliment, ustensile, action, etc.) et détermine les interactions entre ces éléments (exemple : l'unité lexicale « pelée » correspond à une action de façonnage des légumes).

Plusieurs travaux ont déjà exploité l'enrichissement lexical dans leurs approches. Les travaux de (Hamada, 2003) ont étudié le besoin des systèmes de génération ou de l'interprétation de documents textuels, en termes d'informations supplémentaires pour sélectionner des termes spécifiques ou une signification en fonction du contexte. Ce processus doit être soutenu par des informations lexicales supplémentaires en termes de caractéristiques exprimant les relations entre les éléments lexicaux tels que (synonymie, polygamie, antonymie, etc.).

(Nouvel *et al.*, 2013) proposent une approche pour la reconnaissance d'entités nommées. Elle intègre des techniques de fouille séquentielle hiérarchique pour la détection des entités nommées. Le système adopte une démarche centrée sur les données pour extraire des motifs symboliques. Il repose par ailleurs sur une stratégie originale qui consiste à rechercher séparément le début et la fin des entités.

(Mallat *et al.*, 2013) proposent une méthode d'enrichissement lexical des requêtes en arabe afin d'améliorer les performances des systèmes de recherche d'information SRI.

Cette méthode permet deux types d'enrichissement : linguistique et contextuel. Le premier est basé sur l'analyse linguistique de la lemmatisation, l'analyse morphologique, syntaxique et sémantique, dont le but est de générer une liste de listes descriptives. Cette liste contient un ensemble de lexiques linguistiques attribué à chaque terme significatif de la requête. Le deuxième enrichissement consiste à intégrer des informations contextuelles issues des documents du corpus. Il est basé sur une analyse statistique.

(Amato *et al.*, 2017) ont proposé une approche d'extraction et de gestion de données significatives, à partir des occurrences de texte dans des corpus de diagnostic médicaux. L'approche morphosémantique est capable d'accéder à la dimension sémantique des mots et des phrases qu'elle contient.

L'approche choisie consiste à exploiter deux axes pour l'analyse des entités (voir figure 6.10) :

— ontologique : une unité, peut-être graduellement généralisée (par exemple, tomate-> fruit_légume-> légume-> aliment) et des informations lui seront associées (axe paradigmatique) ;

— structurelle : une analyse peut avoir à prendre en considération plusieurs unités contiguës (par exemple, tomates (NOM), peler (VER), etc.)(axe syntagmatique).

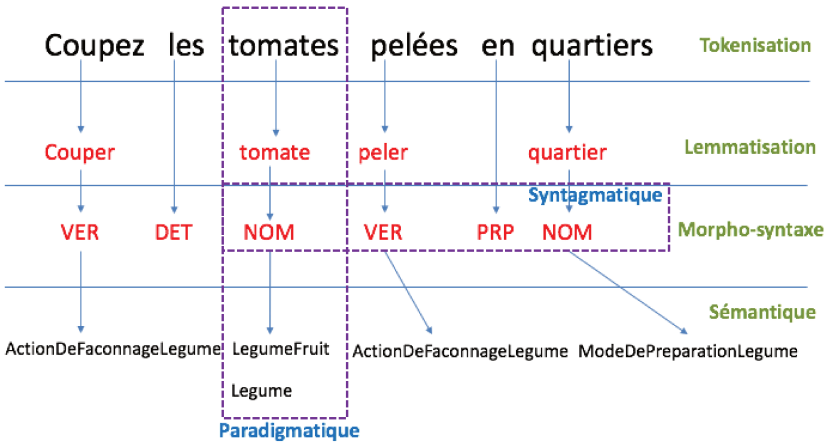


Figure 6.10 – Axes de recherche pour l'enrichissement sémantique des tokens

En définissant la manière dont les structures abstraites sont interprétées, le lexique joue un rôle crucial dans notre proposition. Le résultat de cette analyse peut-être indéterminée et conduire à envisager plusieurs hypothèses. Considérons l'exemple de l'instruction « grattez les moules ». Tout d'abord, on peut noter qu'il est possible que deux interprétations partagent un même vocabulaire abstrait. Même si nous nous situons dans un unique contexte (culinaire), une ambiguïté demeure sur le terme « moule ». En effet, il peut faire référence à un « ustensile » ou à un « aliment ». Pour résoudre ce problème, une classe est affectée à chaque token. Cette représentation par affectation de classes aux termes permet d'utiliser des modèles estimant la probabilité qu'une classe soit affectée à un token et non à un autre (par exemple, Moule est affecté à Ustensile ou Aliment avec une certaine probabilité en fonction du contexte). Il existe différents modèles adaptés à ce type de problème, parmi ces derniers figurent les modèles bayésiens et le clustering (Miller *et al.*, 2004).

Techniquement, nous avons initié la première étape de la méthodologie par une segmentation de la recette qui produit une liste d'ingrédients et une liste d'instructions. Chaque ingrédient est représenté selon le patron « produit/quantité/unité ». Par

exemple , « 600g cuisses de poulet » devient « cuisse de poulet/600/g ». Le « motif » correspondant au « produit » est lemmatisé de manière à favoriser la reconnaissance des termes (suppression des formes plurielles et féminisées par exemple, cuisses de poulet -> cuisse de poulet). La recherche lexicale dans la RTO du motif lemmatisé « cuisse de poulet » permet d'accéder à la hiérarchie des concepts associés à ce motif (voir figure 6.11) ainsi qu'à d'autres informations, par exemple, le module auquel il appartient (Aliment), son concept dans l'ontologie (CuisseDePoulet), etc.

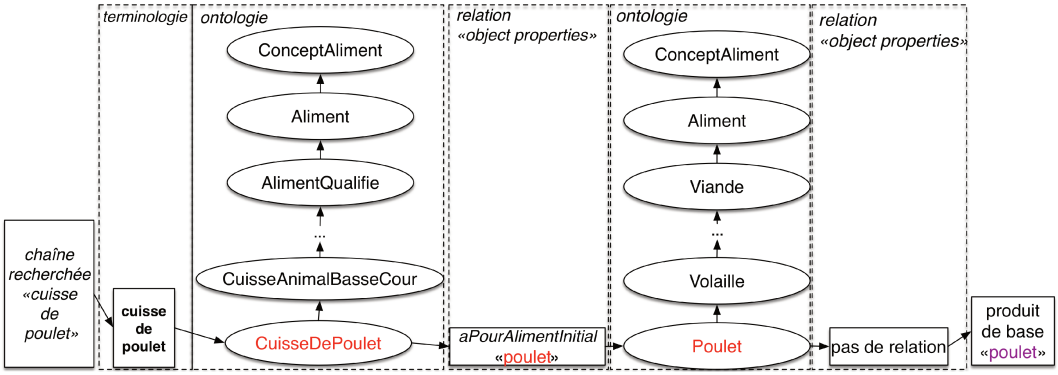


FIGURE 6.11 – Recherche de cuisse de poulet dans la RTO

Après avoir identifié le « concept produit » dans l'ontologie, une recherche permettant de suivre les relations transversales exprimées sous forme d'« ObjectProperty » dans l'ontologie est exécutée. Cette ontologie décrit des formes plus complexes d'aliments telles que des aliments modifiés (beurre salé), transformés (saumon fumé) ou particularisés (cuisse de poulet). Dans ce travail, les relations utiles sont : « aPourAlimentInitial » et « aPourAlimentOrigine ». Par exemple (voir figure 6.11), si nous suivons la relation « aPourAlimentInitial » qui est associée au concept « CuisseDePoulet », le produit de base pour ce concept est « Poulet ». Ce qui nous permet d'identifier le produit de base pour le produit « cuisse de poulet » qui est « poulet ». Il en va de même pour la relation « aPourAlimentOrigine ». Cette approche nous permet d'identifier le produit de base correspondant à chaque produit dans la liste d'ingrédients. Les nouvelles connaissances acquises sont liées au produit.

Après avoir traité les ingrédients, le second traitement porte sur la préparation. Il débute par le découpage de la préparation en instructions (exemple d'instruction dans la préparation : faire dorer les cuisses de poulet dans l'huile). Ces instructions seront ensuite segmentées pour obtenir une liste de motifs lexicaux (exemple : faire, dorer, poulet, etc.). Chaque motif lexical sera recherché dans la RTO et les résultats identifiés seront renvoyés sous la forme « Module : identifiant du concept » (exemple, « dorer » donne « PREPARATION : DORER »). Deux listes sont créées à partir des résultats obtenus : une première liste regroupant les concepts identifiés dans le module ALIMENT (dans notre exemple : Poulet, Huile) et une seconde avec ceux reconnus dans le module PREPARATION (dans notre exemple : Dorer, Cuisse). Dans le cas où le « produit de base » appartient à la première liste, l'ensemble des concepts de la seconde liste correspondant à la même instruction sont associés à ce « produit de base ». Dans notre exemple, la recherche du « produit de base » « poulet » dans la première liste donne un résultat positif, ce qui implique que le produit « cuisse de poulet » reçoit l'ensemble des items de la seconde liste à savoir « Dorer et Cuisse ».

Pour identifier l'état du produit dans la recette et la catégorie à laquelle il appartient (viande, légume, etc.), la recherche réalisée dans le lexique est sous forme d'interroga-

Aliment	Energie (kcal)	Protéines (g)	lipides (g)	fibres (g)
tomate	20.1	0.91	0.33	1.50
tomate cuite	30.2	1.41	0.50	2.8

TABLE 6.2 – Valeurs nutritionnelles de la Tomate fraîche et cuite

tion : « Identifiant concept » est-il un sous-type du « concept générique » ?. Le concept générique pour l'état du produit est « ActionDeCuisson » et celui de la catégorie du produit est « Fruit », « Légume », « Viande », etc. Pour identifier l'état du produit dans la recette, l'interrogation est réalisée à partir des items de la seconde liste associés au produit. Dans notre exemple, « Dorer » et « Cuisse », l'interrogation sera de type « Dorer » est-il sous-type du concept générique « ActionDeCuisson » ?. Si la réponse est positive alors l'état de mon aliment est « cuit », dans le cas contraire il est « cru ». Cette distinction de l'état de l'aliment est cruciale, car la cuisson affecte profondément la qualité nutritionnelle de l'aliment, d'autant plus qu'il existe différentes méthodes de cuisson. Par exemple, une cuisson au four fait perdre plus d'eau par rapport à une cuisson à l'eau. Le tableau 5.1 présente les valeurs nutritionnelles de l'aliment « tomate » selon qu'il soit cuit ou non.

Pour la catégorie du produit, l'interrogation est de type « identifiant concept produit » est-il sous-type du concept générique « Fruit » ?, « Légume » ?, « Viande » ?, etc. Tant que la réponse n'est pas positive l'interrogation se poursuit jusqu'à l'identification de la catégorie. Le résultat obtenu concernant la catégorie de l'exemple « Poulet » est « Viande ».

6.4.4 Génération du fichier de calcul

La deuxième étape de l'algorithme consiste à produire un fichier de calcul qui est structuré sous forme d'un patron « identifiant produit/catégorie/poids ». Cet identifiant du produit correspond à celui figurant dans la base de données nutritionnelle. En partant de la liste des produits enrichis dans la première étape, une chaîne de caractères est construite selon le patron « produit de base/partie du produit/action de cuisson/autre modification ». L'identification de chacun des termes du patron dans la liste des termes du « produit enrichi », passe par l'interrogation de l'ontologie. La démarche est la même pour chacun des termes du patron. Par exemple, pour la « partie du produit » l'« identifiant concept » à identifier est « PartieUtileDe ». L'interrogation de l'ontologie : « identifiant concept » est-il sous-concept générique de « AlimentPartieUtile » ? (1) si la réponse est positive l'identifiant concept correspond à la « partie du produit » ; (2) si après avoir testé tous les termes du produit enrichi aucun résultat n'a été identifié la « partie du produit » prend la valeur par défaut « entier ».

Pour identifier l'« action de cuisson » on vérifie l'état du produit identifié dans la première étape (par exemple tomate cuite). Si l'état du produit est « cuit » alors la valeur de « action de cuisson » est celle de l'« identifiant concept » avec lequel on a identifié l'état de l'aliment dans la première étape (par exemple dans le cas de « cuisse de poulet rôti », l'action de cuisson est « Rotir »). Dans le cas contraire, la valeur sera celle de l'état du produit déjà identifié dans la première étape (cru).

Une requête SQL sur la base de données est ensuite formulée, pour récupérer toutes les lignes dans lesquelles apparaît la chaîne « produit de base » (voir annexe B). Le résultat de cette requête constitue la liste de candidats potentiels pour le produit. Ensuite, cette liste est réduite jusqu'à l'identification du produit. La solution consiste

à utiliser un comparateur pour spécifier l'ordre du tri. Chaque fois que la méthode de tri doit comparer deux éléments, elle utilise un token du patron précédemment construit (produit de base/partie du produit/action de cuisson/autre modification). Par exemple, le premier comparateur sera « produit de base », le second « partie du produit », etc.

Prenons à titre d'exemple le patron construit pour l'ingrédient « tomate 450 g» qui est « tomate/cuit/ ». La requête SQL pour récupérer les lignes contenant le produit « tomate » donne la liste suivante :

$$\begin{array}{l} \textit{tomate}\ \{id : 733\} \\ \textit{tomate à la provençale}\ \{id : 1325\} \\ \textit{tomate cerise}\ \{id : 1261\} \\ \textit{tomate cuite}\ \{id : 3306\} \\ \textit{tomate entière pelée au jus}\ \{id : 820\} \\ \textit{tomate farcie à la chair à saucisse}\ \{id : 959\} \\ \textit{tomate séchée à l'huile, égouttée}\ \{id : 4962\} \end{array} \tag{6.2}$$

Le premier filtre consiste à garder que les éléments de cette liste contenant le second token du patron « tomate/cuit/ » qui est « cuit ». Le résultat obtenu est le suivant :

$$\textit{tomate cuite}\ \{id : 3306\} \tag{6.3}$$

Ainsi, les valeurs nutritionnelles qui seront utilisées pour le calcul sont celles correspondant à l'ingrédient « tomate cuite » portant l'identifiant « 3306 ».

Les unités de chaque produit sont converties en gramme, grâce à une table de conversion fournie par l'EREN, pour obtenir les nouvelles quantités, car les valeurs nutritionnelles dans les tables de composition sont exprimées pour 100g. De plus, les formules de calcul des valeurs nutritionnelles sont elles aussi fondées sur des poids exprimés en gramme. La procédure est la suivante : (1) si l'unité du « produit de base » est déjà en gramme alors la valeur du poids ne change pas ; (2) si l'unité du « produit de base » n'est pas exprimée en gramme, une table de conversion de calibre du produit intervient. Par exemple, l'ingrédient « poivron/3/unité ». Le poids moyen d'un poivron est de 50g, le poids en gramme du poivron dans la recette vaut 3x50=150g. A l'issue de ce traitement, chaque produit sera qualifié avec un poids en gramme. Enfin, le fichier de calcul est établi et est utilisé pour le calcul du score.

6.4.5 Calcul et attribution du score

Le principe d'un étiquetage nutritionnel simplifié (Nutri-Score) a été suggéré et adopté en France dans le cadre du Programme national nutrition santé. L'objectif du Nutri-Score est d'aider le consommateur à savoir si les chiffres indiqués sur les étiquetages (sucres, graisses ou sel par exemple) sont raisonnables ou non. Ceci nous a conduit à construire notre méthodologie sur la base de ce système en l'étendant à des recettes de cuisine.

Différentes méthodes de calcul de la valeur nutritionnelle des recettes existent (Charrondiere *et al.*, 2011). Ces méthodes se différencient en fonction de l'application ou non des facteurs de rendement (changement du poids des aliments ou des recettes dû à la cuisson) et de rétention des nutriments (vitamines et minéraux après cuisson, stockage, préparation, réchauffement, etc.) et le niveau auquel intervient ce calcul.

Le calcul des valeurs nutritionnelles est effectué en automatisant la méthode mixte du calcul de la valeur nutritionnelle d'une recette. Le calcul du facteur de rendement est obtenu à partir du poids de la recette une fois crue et une fois cuite ; le facteur de rendement est appliqué au niveau de la recette et les facteurs de rétention des nutriments au niveau des ingrédients (Ducrot, 2016). Cette démarche a été validée par les diététiciens de l'Equipe de Recherche en Epidémiologie Nutritionnelle. Il se déroule comme suit : (1) calcul du poids net du plat final et de celui de chaque ingrédient en appliquant les coefficients de rendement qui sont dépendants de la catégorie du produit (fruit, légume, viande, etc.) ainsi que son état dans le plat final (cuit ou cru) ; (2) calcul du poids d'une portion du plat par individu ; (3) calcul du pourcentage de fruits et légumes (%FL) en appliquant la formule suivante :

$$\%FL = \frac{poids\ FL\ +\ (2\ *\ poids\ FL\ seche)}{poids\ FL\ +\ (2\ *\ poids\ FL\ seche)\ +\ poids\ autres\ ingrédients} \tag{6.4}$$

(4) récupération des valeurs nutritionnelles des éléments nutritifs suivants : protéines, énergie, graisses saturées, sucres, sodium et fibres pour chaque « produit » dans la table de composition des aliments Nutrinet avec une requête SQL ; (5) convertir les valeurs de chaque produit par rapport au poids réel et non pour 100g comme c'est le cas dans la base de données ; (6) récupération des coefficients des parties comestibles pour chaque produit avec une requête SQL ; (7) application des coefficients des parties comestibles pour calculer les valeurs nutritionnelles seulement pour ce qui est comestible dans le produit. Par exemple, le pourcentage comestible d'une « cuisse de poulet » est de seulement 60% ; (8) calcul des valeurs nutritionnelles de chaque élément (VEN) dans le plat final avec la formule suivante :

$$VEN\ A\ du\ plat = \sum_i Valeur\ (partie\ comestible\ )\ de\ chaque\ produit \tag{6.5}$$

(9) conversion des valeurs nutritionnelles de chaque élément nutritif pour 100g de plat.

L'attribution du score est réalisée en appliquant le système de profilage utilisé par la FSA. Le calcul du score est fondé sur la teneur au 100g de 4 éléments négatifs (N) : énergie, sucres simples, acides gras saturés et sel. Des points de 1 à 10 sont attribués à chaque élément en fonction des quantités présentes. La composante N peut donc aller de 0 (le plus favorable) à 40 (le moins favorable). Ce score N peut éventuellement être minimisé par la soustraction d'éléments nutritionnels considérés comme positifs (P) : fruits ou légumes (et noix), fibres et protéines. Des points allant de 1 à 5 sont attribués. La note P va donc de 15 (plus favorable) à 0 (moins favorable). L'équation se traduit sous cette forme :

$$\begin{array}{l} Si\ les\ points\ N\ <11,\ alors\ score = points\ N - points\ P; \\ Si\ les\ points\ N\ >11\ : \\ Si\ les\ points\ des\ FL = 5,\ alors \\ score = points\ N - points\ P; \\ Si\ les\ points\ des\ FL < 5,\ alors \\ score = points\ N - points\ (points\ fibres) + (points\ FL); \end{array} \tag{6.6}$$

Les notes théoriques du score vont donc de -15 (le plus favorable) à +40 (le plus défavorable). Des couleurs et des lettres sont attribuées en fonction du score final : A/vert (de -15 à -2) ; B/jaune (de -1 à 3) ; C/orange E/rouge (de 17 à 40) (Hercberg, 2013).

6.4.6 Traduction du score sur l'échelle graphique

Pour chaque recette, le score sera visualisé en utilisant l'échelle nutritionnelle du PNNS proposé par Serge Hercberg dans le cadre du rapport sur les mesures concernant la prévention nutritionnelle. Cette échelle se présente sous forme de 5 disques de couleurs différentes sur une échelle allant du vert au rouge, couplée à une lettre correspondante sur l'échelle des « notes » d'école de A à E pour plus de compréhension. La taille des disques dépend du score atteint par la recette (voir figure 6.12).

L'influence des systèmes de représentation graphique des valeurs nutritionnelles des aliments sur la capacité des individus à classer les produits en fonction de leur qualité nutritionnelle a été évaluée par l'étude (Ducrot *et al.*, 2014). Les quatre systèmes comparés sont : (1) le logo Pick the Tick ; (2) les Multiple Traffic Lights (MTL) ; (3) les Guideline Daily Amounts (GDA) et (4) l'échelle de qualité nutritionnelle du PNNS (échelle de qualité nutritionnelle du PNNS). Les résultats obtenus ont démontré que les systèmes d'information nutritionnelle augmentent de façon significative la capacité des individus à classer les produits en fonction de leur qualité nutritionnelle. Ainsi, les résultats obtenus sont en faveur du PNNS qui donne des meilleurs résultats indépendamment de l'âge, du sexe et de la catégorie socio-économique .

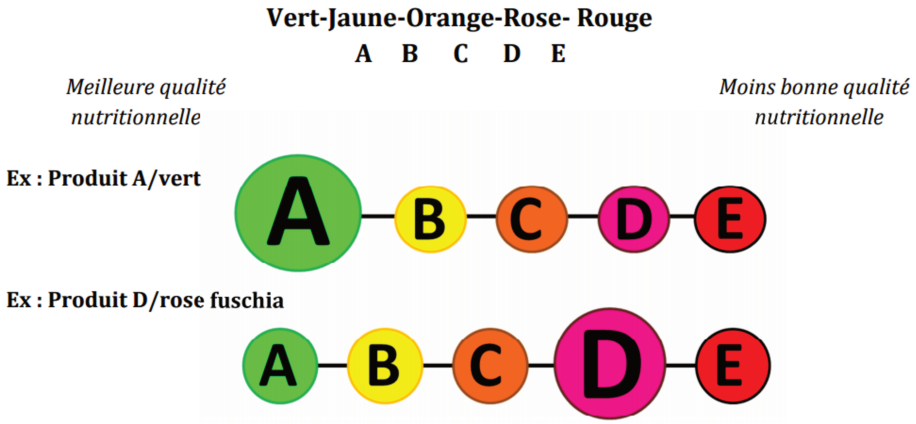


Figure 6.12 – Echelle de qualité nutritionnelle du PNNS

6.5 Evaluation des résultats

L'approche proposée a été évaluée sur une dizaine de recettes issues du corpus des 45 recettes déjà évaluées manuellement par les diététiciens de l'EREN. Pour ce premier échantillon, la comparaison a été effectuée entre trois outils disponibles sur internet (« nutryaccess » [18], « myfitnesspal » [19], « monmenu » [20].) et notre outil.

Le tableau 6.3 présente les résultats obtenus pour les six premières recettes de l'échantillon test. La première colonne du tableau correspond aux valeurs nutritionnelles des recettes calculées par les diététiciennes de l'EREN, la deuxième aux valeurs obtenues par notre outil de calcul et la troisième à celles obtenues par l'outil « monmenu ».

A partir du tableau 6.3, nous pouvons faire une comparaison des résultats produits par notre outil avec les valeurs obtenues pour la recette « ratatouille » établie par les

18. http://www.nutryaccess.com/
19. https://www.myfitnesspal.fr/
20. http://www.monmenu.fr/

	Enérgie Kcal/100g			Protéines g/100g			Acide gras saturé g/100g		
	EREN	Outil	Monmenu	EREN	Outil	Monmenu	EREN	Outil	Monmenu
Blanquette de veau	149	171	127	11	13	9	5	6	8
crêpes	225	237	435	6	7	9	1	1	19
Gâteau aux pommes	237	254	300	3	3	6	1	2	4
Moelleux au chocolat	458	413	485	5	5	4	6	17	24
Ratatouille	83	100	33	1	1	1	1	1	2
Salade niçoise	108	119	71	8	12	8	1	2	3
	Sucres simples g/100g			Sodium Mg/100g			Fibres g/100g		
	EREN	Outil	Monmenu	EREN	Outil	Monmenu	EREN	Outil	Monmenu
Blanquette de veau	1	1	2	231	118	/	1	1	1
crêpes	4	4	57	153	61	/	1	1	/
Gâteau aux pommes	26	24	35	77	29	/	1	1	/
Moelleux au chocolat	40	30	63	166	71	/	1	2	/
Ratatouille	1	1	2	202	211	/	2	2	/
Salade niçoise	1	1	2	517	444	/	2	1	/

TABLE 6.3 – Résultats pour les six premières recettes

diététiciens de l'EREN d'une part et par un système équivalent à notre outil d'autre part (voir figure 6.13).

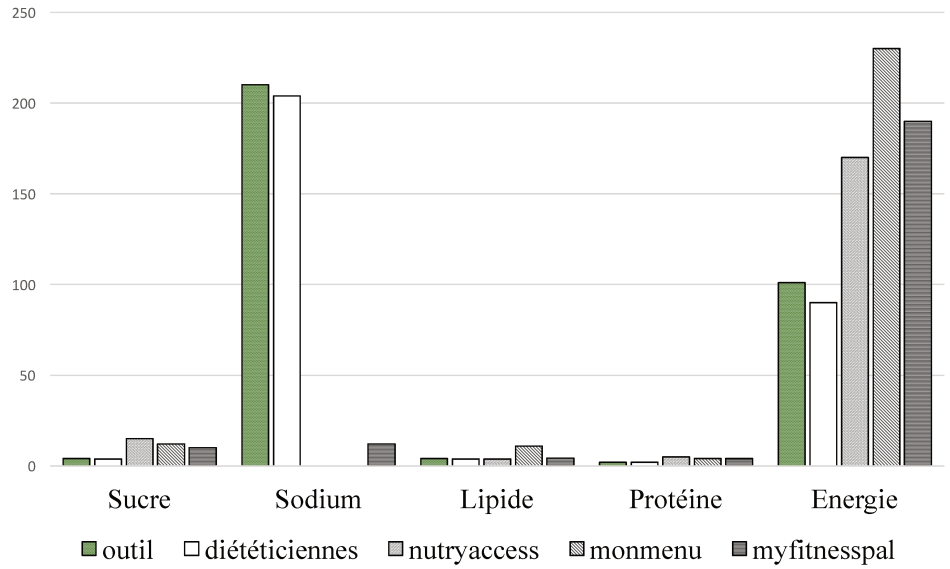


FIGURE 6.13 – Comparaison sur la recette ratatouille

Nous constatons sur la figure 6.13 que les valeurs obtenues par notre outil sont proches de celle obtenue par les diététiciennes de l'EREN, contrairement à celle des trois autres outils. Cette différence significative entre les valeurs obtenues peut s'expliquer de plusieurs façons :

— la première concerne les tables de composition des aliments utilisées pour le calcul des VNs de la recette. Les trois outils comparés ne fournissent aucune information sur la nature et la provenance des données utilisées alors que nous utilisons la table Nutrinet développée par l'EREN qui répond à des critères de qualité supérieurs.

— La seconde porte sur le traitement de la recette. D'un côté, les deux premiers outils comparés traitent la recette d'une manière semi-automatique. Chaque ingrédient de la recette est choisi manuellement par l'utilisateur dans une liste d'ingrédients (le choix dépend des capacités culinaires de l'utilisateur). Par exemple, dans une liste d'ingrédients figure boeuf, une personne peut choisir « côte de boeuf » qui a pour valeur lipidique 21g/100g et une autre personne choisira « entrecôte de boeuf » dont la valeur lipidique est de 11g/100g. De l'autre, notre outil et celui de « monmenu » traitent automatiquement la recette. Ce traitement porte sur la liste d'ingrédients de manière à apparier chaque ingrédient avec un ingrédient présent dans la table de composition. Pour gagner en précision dans l'appariement, notre outil traite non seulement la liste d'ingrédients, mais l'ensemble de la recette (préparation) ce qui nous permet d'être plus précis que l'outil « monmenu ». Par exemple, pour l'ingrédient « cuisse de poulet » qui sera rôtie dans le plat, avec l'outil « monmenu » la valeur de lipide utilisé pour le calcul sera de 7g/100g. Grâce au traitement que nous avons effectué sur la préparation, nous avons pu déterminer l'état de l'ingrédient (rôtie) dans le plat. Ceci nous a permis d'apparier « cuisse de poulet » avec « cuisse de poulet rôtie » dans la BD Nutrinet. La valeur lipidique de cette dernière est alors de 18g/100g.
L'autre point commun aux trois outils comparés correspond à la méthode employée pour calculer les VNs d'une recette. Cette méthode consiste simplement à

additionner les VNs de chaque ingrédient. Cette méthode est limitée puisqu'elle ne prend pas en compte ni l'état du produit dans le plat final (cuit, cru) ni la variation due au processus de préparation du plat. Or l'état de l'aliment dans le plat final est essentiel pour obtenir des valeurs correctes. Par exemple, tomate dans la liste d'ingrédients qui sera cuite dans le plat, aura pour valeur d'énergie 20kcal/100g pour « monmenu », mais 30,2 Kcal/100g pour notre outil. Dans le cas où l'aliment « tomate » est saisi avec le paramètre cuit (« tomate cuite » au lieu de « tomate ») dans l'outil « monmenu », le système appariera « tomate cuite » avec « flageolet cuit » dont la valeur d'énergie correspond à 74Kcal.

— La troisième porte sur la méthode de conversion appliquée pour passer d'une unité à l'autre (par exemple, le poids moyen d'une pomme en gramme). Aucune information n'est disponible sur la méthode de conversion pour les trois outils. Les unités ménagères diffèrent selon les pays parce que les dimensions des aliments et les mesures ménagères sont différentes. La copie des unités ménagères à partir de celles d'autres pays doit être évitée, car elle pourrait introduire des erreurs importantes pour les teneurs en nutriments d'une recette ou pour d'autres calculs.

— Enfin une autre faiblesse observée au niveau des trois outils comparés correspond aux informations nutritionnelles fournies à l'utilisateur concernant le plat préparé. Par exemple dans la plupart des cas aucune information n'est fournie pour la valeur du sodium, le pourcentage des fruits et légumes dans le plat, etc. Notre méthodologie fournit en plus de ces informations une note traduisant la qualité nutritionnelle globale du plat consommé.

D'un point de vue méthodologique, la démarche proposée doit encore être testée de manière plus approfondie. Une évaluation des différentes étapes de l'algorithme semble essentielle pour quantifier chacune de ces dernières. Par exemple, le taux de reconnaissance des produits composant les ingrédients, l'impact de la reconnaissance sur la valeur du score final de la recette, le taux d'affectation de la bonne méthode de cuisson, etc.

En ce qui concerne la précision de l'appariement entre les produits de la recette et leurs entrées dans la table de composition les résultats obtenus sont prometteurs et s'avèrent même concluants dans certains cas. Une normalisation des libellés de la table de composition Nutrinet s'avère primordiale pour un meilleur appariement. Nous pourrons pour cela nous appuyer sur un système respectant un standard de référence comme LANGUAL qui est un thesaurus multilingue à facettes créé pour décrire des produits alimentaires d'une façon systématique. Il est fondé sur le principe que n'importe quelle nourriture (ou produit alimentaire) peut être systématiquement décrite par une combinaison de caractéristiques (aussi appelé descripteurs). Ces caractéristiques peuvent être catégorisées dans des points de vue appelés facettes exemple (type de produit, origine biologique, méthodes de cuisson, méthodes de conservation, etc.). Chaque descripteur est identifié par un code unique (Møller et Ireland, 2018 ;Ireland et Møller, 2013). Par exemple, l'aliment « cuisse de poulet rôti (haut de cuisse) », après analyse sera décrit : l'« aliment principal » est le Poulet (B1457) suivi de la facette « Partie utile » (Z0136), « Mode de Cuisson » (H0391) et « Précisions Diverses » (Z0160).

6.6 Conclusion

L'alimentation restant une activité quotidienne, les habitudes alimentaires ont beaucoup changé au cours des dernières années. Ces profondes modifications comportent,

sur le plan nutritionnel et sur le plan de la santé, des aspects positifs et d'autres négatifs, des avantages et des inconvénients pour la santé. C'est pourquoi il est essentiel de repenser et réinventer les notions de simplification et de l'éducation pour la santé, au quotidien, de la vie des utilisateurs. L'objectif du travail présenté dans ce chapitre est de construire un module de qualification nutritionnelle des recettes de cuisine qui va s'intégrer dans un écosystème de prévention du risque cardiovasculaire, pour fournir aux utilisateurs, des contenus et des outils pour mieux les renseigner et les guider dans l'adoption d'un mode de vie saine.

Notre méthode permet de calculer un score nutritionnel associé à une recette. Elle devra être testée d'une manière plus approfondie et sur un plus large corpus de recettes pour avoir une idée de la qualité des résultats. Ce travail souligne les difficultés rencontrées au cours du traitement des données non structurées, ce qui est souvent le cas dans les applications du monde réel. Dans notre cas, il s'agit de ressources textuelles nécessitant des phases préparatoires de segmentation et de lemmatisation. D'autres difficultés se situent au niveau de l'appariement entre les unités (mots, liste de mots, etc.) de la recette et les unités dans la base de données, même dans le cas où il est réalisé manuellement. Cela est dû à une faible normalisation des termes utilisés dans les tables de composition. Lorsqu'en outre la résolution du problème initial doit faire appel à des connaissances implicites, le recours à des ressources sémantiques termino-ontologiques s'avère fructueux, car il permet une meilleure précision lors des traitements ultérieurs purement algorithmiques. Enfin, ce travail ouvre des perspectives de recherche intéressantes en formalisant les connaissances du domaine du suivi nutritionnel. Par exemple, la prise en compte des qualifications sensorielles des aliments permettrait d'envisager des systèmes de suggestions personnalisées pour des personnes souffrant de maladies chroniques en tenant compte de leurs goûts pour induire une meilleure acceptabilité et observance de leurs régimes alimentaires.

Conclusion et perspectives

Arrivé au terme de cette thèse, il est désormais temps de dresser le bilan des principaux acquis et de proposer des perspectives.

Nous venons de présenter nos travaux de recherche : la création d'un ecosystème en proposant un ensemble de services et de méthodes pour la gestion efficace de la prévention du risque. Cette démarche, a été appliquée à la prévention du risque cardiovasculaire. Nos travaux s'intégrent dans l'approche du web sémantique, en s'appuiyant sur les théories, les méthodes et les techniques en ingéniérie des connaissances.

Pour répondre à la problématique énoncée en introduction, nous soutenons l'idée suivante : **les approches de visualisation sont adaptées pour : (i) améliorer la compréhension des interactions entre facteurs de risque quel que soit le domaine ; (ii) faciliter l'accès aux connaissances et aider à créer de nouvelles connaissances. La représentation graphique fait partie des systèmes de signes que l'homme a construit pour retenir, comprendre et communiquer ; (iii) faciliter le travail collaboratif pour la prise de décision entre plusieurs apprenants (par exemple médecin et patient).**

Pour entamer cette thèse, nous avons présenté les notions qui constituent les fondements de notre recherche. A partir de l'étude des modèles et des méthodes de prévention, nous avons construit notre méthodologie de travail. L'analyse de la représentation et de l'organisation de la connaissance nous a permis d'identifier les éléments de notre modèle. Grâce aux langages, recommandations et projets existants dans la littérature, nous avons construit nos modèles et développé des services d'accès aux connaissances sur les interactions entre les facteurs de risque.

La méthodologie que nous avons mise en place pour représenter les interactions entre les facteurs de risque repose sur une approche issue de l'ingénierie des connaissances. Dans un premier temps, cette approche nous a permis d'extraire les connaissances implicites sur les interactions entre les facteurs de risque cardiovasculaire représenté dans des tableaux statistiques (traité dans le chapitre 2). Lors de cette étape , nous avons pu mener en parallèle des expérimentations dans d'autres domaines (domaine financier) pour évaluer l'extraction de connaissances et une comparaison des différents modèles obtenus à partir de différentes études statistiques pour évaluer la pertinence des connaissances extraites. Dans un deuxième temps, nous avons représenté les connaissances extraites sur les facteurs de risque sous forme d'un modèle conceptuel (traité dans le chapitre 3). Ce choix était motivé par le besoin de construire un modèle générique sur les interactions entre facteurs de risque cardiovasculaire. En effet, nous devons intégrer les facteurs de risque est leurs interactions dans le processus d'évaluation du risque. Dans un troisième et dernier temps, nous avons mis en place une interface de visualisation sous forme d'un service web permettant au médecin d'évaluer le risque

cardiovasculaire d'un utilisateur (traité dans le chapitre 4). Cette approche a été intégrée dans l'écosystème apprenant et construite de manière modulaire (traité dans le chapitre 5). Dans ce système, différentes recommandations sont élaborées en prenant en compte le risque présenté par l'utilisateur. Parmi ces modules un service web a été implémenté pour suggéré une alimentation saine aux utilisateurs. Il a donc été nécessaire d'implémenter un algorithme de qualification nutritionnelle des recettes (traité dans le chapitre 6).

Contributions méthodologiques. L'originalité de nos travaux réside dans leur caractère interdisciplinaire et dans l'utilisation conjointe d'approches issues de l'ingénierie des connaissances et du web sémantique. Notre approche tend à : (i) rendre générique la modélisation des interactions entre des facteurs de risque. L'approche proposée permet d'extraire les connaissances à partir d'études statistiques publiées sous forme de tableaux de formats différents ; (ii) permet de réaliser une qualification nutritionnelle des recettes de cuisine avec une traduction visuelle de ces valeurs nutritionnelles selon le Nutri-Score.

Contribution techniques. L'implémentation et l'utilisation de l'outil MCVGraphViz ont permis le développement d'une nouvelle approche dynamique d'évaluation du risque dans différents domaines en traitant des interactions entre facteurs de risque. L'outil MCVGraphViz est intégré dans l'écosystème apprenant développé dans cette thèse.

Le service de qualification nutritionnelle qui fait partie de l'écosystème permet de calculer les valeurs nutritionnelles des recettes de cuisine et d'attribuer une note représentée avec le logo du Nutri-Score. Il aide ainsi l'utilisateur dans le choix d'une alimentation saine. Ce service est réalisé on ce basant sur une démarche experte, celles des nutritionnistes de l'EREN.

Perspectives de recherche. Les contributions obtenus font émerger de nouvelles pistes de recherche à explorer.

— **Pour l'évaluation du risque cardiovasculaire** : nous avons montré que l'évaluation du risque cardiovasculaire peut être améliorée en utilisant les standards du web sémantique. Il serait important de décrire chaque facteurs de risque et de mieux comprendre les liens entre ces facteurs qui peuvent être de nature différentes. Le passage à une ontologie semble nécessaire pour mieux décrire certaines relations entre les facteurs.

— **Pour la qualification nutritionnelle des recettes de cuisine** : Il serait intéressant de compléter les expérimentations réalisées sur les recettes de cuisine. L'objectif serait alors d'évaluer cette approche sur l'ensemble du corpus des recettes. Plus précisément, il s'agirait de vérifier si les résultats obtenus sont robustes et représentatifs. Afin d'améliorer la précision du calcul, il faudrait réaliser des expérimentations sur l'ordre du patron utilisé lors de phase d'appariement des ingrédients avec les aliments de la table de composition (annexe A). Il s'agirait d'examiner l'impact de l'ordre de chaque élément du motif construit sur les valeurs nutritionnelles de la recette. Diverses questions sont soulevées : la partie de l'aliment est-elle plus importante que l'action de cuisson ? Y'a-t-il d'autres paramètres à prendre en compte ? Quel est le paramètre par défaut lorsque ce dernier n'est pas identifié par le traitement automatique ? Il serait intéressant de prendre en compte l'ordre des tokens dans les instructions pour corriger certaines ambiguïtés. Par exemple, si le concept « ramequin » est identifié dans une instruction, il peut correspondre à un aliment (fromage) ou à un ustensile de cuisine. L'approche serait de lister les termes les plus proches pour identifier le contexte

(par exemple, nous avons identifié « Four », « Mettre » alors dans le contexte ramequin correspond à un ustensile de cuisine).

— **Pour la réutilisation des données** : Une réflexion devrait être menée pour déterminer les enjeux de la réutilisation des ressources identifiées dans le chapitre 1. Il conviendrait d'identifier les données susceptibles d'améliorer l'écosystème construit et leurs contextes d'usage. Il faudrait s'interroger sur les statuts juridiques de ces données et comprendre la manière dont ces données circulent entre les différents acteurs afin de créer une valeur ajoutée à l'information réutilisée.

— **Pour le système de recommandation** : Pour conduire des expérimentations sur les fonctionnalités proposées, nous prévoyons à court terme de déployer l'écosystème sur une architecture client-serveur.

Nous prévoyons d'analyser les retours d'expérience des utilisateurs sur les recommandations qui leur sont proposées. Cette analyse permettra d'évaluer l'intérêt de nos modèles pour la recommandation d'items pertinents et d'évaluer la satisfaction des utilisateurs. Ces retours pourront être exploités par le système de recommandation en vue d'affiner les profils utilisateurs.

L'étude de l'évolution des préférences dans le temps s'inscrit également dans le cadre de nos perspectives de recherche. En effet, les appréciations des utilisateurs ont tendance à évoluer avec le temps. Ainsi, notre objectif est de proposer un système de recommandation capable de détecter le changement du comportement de l'utilisateur et d'adapter dynamiquement les recommandations en fonction des nouveaux besoins de cet utilisateur.

Nous prévoyons également à long terme : (i) d'enrichir les modules existants en proposant de nouvelles fonctionnalités telles que la géolocalisation pour être en mesure de proposer des restaurants en lien avec son profil et son environnement ; (ii) d'intégrer de nouveaux modules pour mieux répondre aux besoins des utilisateurs. Par exemple, un module sensoriel pour mieux compléter la suggestion, un module sur la santé sous forme de jeux sérieux (*serious games*) qui s'adresse à l'utilisateur ou à ses proches (par exemple, le jeu sérieu « Aidant-et-Eve [21] » qui s'adresse aux familles de patients atteints de la maladie d'Alzheimer à un stade léger ainsi qu'aux pathologies assimilées).

— **Pour la visualisation** : La validation de MCVGraphViz peut être poursuivie selon deux axes. Tout d'abord, l'outil peut être testé avec des médecins et des patients. Ceci permettrait de conforter nos hypothèses sur la stratégie mise en place pour évaluer le risque cardiovasculaire. En outre, de nouvelles connaissances peuvent être acquises, soit sur conseil de l'expert (s'il est capable d'expliciter ces relations), soit par analyse des historiques d'exploration dans MCVGraphViz.

Ensuite, il est possible de comparer MCVGraphViz à d'autres outils de visualisation de graphes selon un protocole expérimental donné. Ce protocole peut consister à demander à deux panels d'utilisateurs d'utiliser MCVGraphViz et un autre outil pour exécuter les mêmes tâches d'exploration sur les mêmes données. Par exemple, la tâche serait : (i) de retrouver dans le graphe des facteurs de risque prédisant un facteur choisi et avec quel pourcentage ; (ii) répondre à une question portant sur les données. Une telle expérience permettrait de comparer l'exactitude et le temps de production des réponses des utilisateurs. En outre, MCVGraphViz sera dans l'avenir utilisé pour évaluer le risque cardiovasculaire dans l'écosystème.

21. `http://www.aidant-et-eve.fr/`

Bibliographie

Abidi, S. et H. Chen (2006). « Adaptable Personalized Care Planning via a Semantic Web Framework ». In : *20th Int Conference of the European Federation for Medical Informatics, 2006*. Maastricht, Netherlands (*citée page* 115, 118).

Achananuparp, P. et I. Weber (2016). « Extracting Food Substitutes From Food Diary via Distributional Similarity ». In : *CoRR* abs/1607.08807 (*citée page* 113).

Aciar, S., D. Zhang, S. Simoff et J. Debenham (2007). « Informed Recommender : Basing Recommendations on Consumer Product Reviews ». In : *IEEE Intelligent Systems* 22.3, p. 39-47. doi : 10.1109/mis.2007.55 (*citée page* 109).

Adam, C. T. (1981). « A descriptive definition of primary prevention ». In : *The Journal of Primary Prevention* 2.2, p. 67-79. doi : 10.1007/bf01333744 (*citée page* 15).

Adobe, P. (1985). *PostScript Language Reference Manual*. 1st. Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc. (*citée page* 49).

Adomavicius, G. et A. Tuzhilin (2005). « Toward the next generation of recommender systems : a survey of the state-of-the-art and possible extensions ». In : *IEEE Transactions on Knowledge and Data Engineering* 17.6, p. 734-749. doi : 10.1109/tkde.2005.99 (*citée page* 106).

Adriouch, S., C. Julia, E. Kesse-Guyot, P. Ducrot, S. Péneau, C. Méjean, K. E. Assmann, M. Deschasaux, S. Hercberg, M. Touvier et L. K. Fezeu (2017). « Association between a dietary quality index based on the food standard agency nutrient profiling system and cardiovascular disease risk among French adults ». In : *International Journal of Cardiology* 234, p. 22-27. doi : 10.1016/j.ijcard.2017.02.092 (*citée page* 129).

Adriouch, S., C. Julia, E. Kesse-Guyot, C. Méjean, P. Ducrot, S. Péneau, M. Donnenfeld, M. Deschasaux, M. Menai, S. Hercberg, M. Touvier et L. K. Fezeu (2016). « Prospective association between a dietary quality index based on a nutrient profiling system and cardiovascular disease risk ». In : *European Journal of Preventive Cardiology* 23.15, p. 1669-1676. doi : 10.1177/2047487316640659 (*citée page* 129).

Aggarwal, C. et H. Wang, éd. (2010). *Managing and Mining Graph Data*. T. 40. Advances in Database Systems. Springer. doi : 10.1007/978-1-4419-6045-0 (*citée page* 86).

Ahire, S. B. et H. K. Khanuja (2015). « A Personalized Framework for Health Care Recommendation ». In : *2015 International Conference on Computing Communication Control and Automation*. IEEE. doi : 10.1109/iccubea.2015.92 (*citée page* 118).

Ahmad, R., M. Tanvir-Afzal et M. Abdul-Qadir (2016). « Information Extraction from PDF Sources Based on Rule-Based System Using Integrated Formats ». In :

Semantic Web Challenges. Springer International Publishing, p. 293-308. DOI : 10.1007/978-3-319-46565-4_23 (*citée page* 49).

AHN, J. wook, P. BRUSILOVSKY, J. GRADY, D. HE et S. Y. SYN (2007). « Open user profiles for adaptive news systems ». In : *Proceedings of the 16th international conference on World Wide Web - WWW '07*. ACM Press. DOI : 10.1145/1242572.1242575 (*citée page* 108).

AKTAS, M. K. (2004). « Global Risk Scores and Exercise Testing for Predicting All-Cause Mortality in a Preventive Medicine Program ». In : *JAMA* 292.12, p. 1462. DOI : 10.1001/jama.292.12.1462 (*citée page* 18).

ALANI, H. (2003). « TGVizTab : An ontology visualisation extension for Protégé ». In : *Knowledge Capture (K-Cap'03), Workshop on Visualization Information in Knowledge Engineering* (*citée page* 91).

ALAYA, N., S. B. YAHIA et M. LAMOLLE (2016). « RakSOR : Ranking of Ontology Reasoners Based on Predicted Performances ». In : *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, p. 1076-1083. DOI : 10.1109/ictai.2016.0165 (*citée page* 28).

ALHAJJ, R. et J. ROKNE, éd. (2018). *Encyclopedia of Social Network Analysis and Mining*. Springer New York. DOI : 10.1007/978-1-4939-7131-2 (*citée page* 30, 86).

ALI, S. I., M. B. AMIN, S. KIM et S. LEE (2018). « A Hybrid Framework for a Comprehensive Physical Activity and Diet Recommendation System ». In : *Lecture Notes in Computer Science*. Springer International Publishing, p. 101-109. DOI : 10.1007/978-3-319-94523-1_9 (*citée page* 112).

ALLSOPP, J. (2007). *Microformats : Empowering Your Markup for Web 2.0*. 1. Apress, p. 368. DOI : 10.1007/978-1-4302-0195-3 (*citée page* 132).

AMATO, F., A. MAZZEO, A. ELIA, A. MAISTO et S. PELOSI (2017). « Morphosemantic strategies for the automatic enrichment of Italian lexical databases in the medical domain ». In : *International Journal of Grid and Utility Computing* 8.4, p. 312. DOI : 10.1504/ijguc.2017.088262 (*citée page* 145).

ANDERSSON, K. et M. S. HOSSAIN (2014). « Smart risk assessment systems using belief-rule-based DSS and WSN technologies ». In : *2014 4th International Conference on Wireless Communications, Vehicular Technology, Information Theory and Aerospace & Electronic Systems (VITAE)*. IEEE. DOI : 10.1109/vitae.2014.6934397 (*citée page* 16).

ANTONIOU, G. et F. van HARMELEN (2004). « Web Ontology Language : OWL ». In : *Handbook on Ontologies*. Springer Berlin Heidelberg, p. 67-92. DOI : 10.1007/978-3-540-24750-0_4 (*citée page* 91).

ARIES, S., B. L. BLANC et J.-L. ERMINE (2008). « MASK : une méthode d'ingénierie des connaissances pour l'analyse et la structuration des connaissances ». In : *Management et ingénierie des connaissances, modèles et méthodes*. Traité IC2, Série Management et Gestion des STIC. Hermes sciences, p. 208 (*citée page* 66, 67).

ARNHEIM, R. (1991). « Perception, cognition, and visualization ». In : *J Biocommun* 18.2, p. 2-5 (*citée page* 81).

ARNOLD, P. et E. RAHM (2014). « Extracting Semantic Concept Relations from Wikipedia ». In : *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics*. 26. ACM Press. DOI : 10.1145/2611040.2611079 (*citée page* 44, 48).

Assurance-Maladie (2010). *Prévention des maladies cardiovasculaires : le nouveau programme de l'Assurance Maladie*. Communiqué de presse. Assurance maladie. 16 p. (*citée page* 4, 11).

Atashi, A., N. Nazeri, E. Abbasi, S. Dorri et M. Alijani-Z (2017). « Breast Cancer Risk Assessment Using adaptive neuro-fuzzy inference system (ANFIS) and Subtractive Clustering Algorithm ». In : *Multidisciplinary Cancer Investigation* 1.2, p. 20-26. doi : 10.21859/mci-01029 (*citée page* 17).

Auer, S. et J. Lehmann (2007). « What Have Innsbruck and Leipzig in Common ? Extracting Semantics from Wiki Content ». In : *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, p. 503-517. doi : 10.1007/978-3-540-72667-8_36 (*citée page* 48).

Aussenac-Gilles, N., J. Charlet et C. Reynaud (2012). « Les enjeux de l'Ingénierie des connaissances ». In : *Information, Interaction, Intelligence - Le point sur le i(3)*. Sous la dir. de P. M. F. Sedes J.M. Ogier. Cepadues (*citée page* 67, 70).

Aussenac-Gilles, N. et M. Kamel (2009). « Ontology Learning by Analyzing XML Document Structure and Content ». In : *KEOD* (*citée page* 47).

Aussenac-Gilles, N., J. P. Krivine et J. Sallantin, éd. (1992). *Acquisition des connaissances*. T. 6, N 2. Revue d'Intelligence Artificielle (RIA). Paris : Hermès (*citée page* 68).

Aussenac-Gilles, N., P. Laublet et C. Reynaud (1996). *Acquisition et ingénierie des connaissances - tendances actuelles*. Editions Cépaduès (*citée page* 69).

Aussenac-Gilles, N. et N. Matta (1994). « Making a method of problem solving explicit with MACAO ». In : *International Journal of Human-Computer Studies* 40.2, p. 193-219. doi : 10.1006/ijhc.1994.1010 (*citée page* 69).

Aussenac-Gilles, N. et P. Séguéla (2000). « Les relations sémantiques : du linguistique au formel ». In : *Cahiers de grammaire*. Université de Toulouse-le-Mirail, p. 175-198 (*citée page* 44).

Azzi, R. (2014). *Conception d'une base de données à partir de la table de composition des aliments Nutrinet*. Rapport de stage M1 Informatique Biomédicale. Université Paris 13, Sorbonne Paris Cité, LIMICS, INSERM, (UMRS 1142), 2014 (*citée page* 141).

Bachimont, B., F. Gandon, G. Poupeau, B. Vatant, R. Troncy, F. Gandon, S. Pouyllau, R. Martinez, M. Battisti et M. Zacklad (2011). « Enjeux et technologies : des données au sens ». In : *Documentaliste-Sciences de l'Information* 48.4, p. 24. doi : 10.3917/docsi.484.0024 (*citée page* 22).

De-Backer, G. (2003). « European guidelines on cardiovascular disease prevention in clinical practice Third Joint Task Force of European and other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of eight societies and by invited experts) ». In : *European Heart Journal* 24.17, p. 1601-1610. doi : 10.1016/s0195-668x(03)00347-6 (*citée page* 17).

Baeza-Yates, R. et B. Ribeiro-Neto (1999). *Modern Information Retrieval*. Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc. (*citée page* 107).

Barton, A., A. Rosier, A. Burgun et J.-F. Ethier (2014). « The Cardiovascular Disease Ontology ». In : *Frontiers in Artificial Intelligence and Applications* 267.Formal Ontology in Information Systems, p. 409-414. doi : 10.3233/978-1-61499-438-1-409 (*citée page* 35).

Baumann, L. C. et A. Ylinen (2017). « Prevention : Primary, Secondary, Tertiary ». In : *Encyclopedia of Behavioral Medicine*. Springer New York, p. 1-3. doi : 10.1007/978-1-4614-6439-6_135-2 (*citée page* 15).

Belleau, F., M.-A. Nolin, N. Tourigny, P. Rigault et J. Morissette (2008). « Bio2RDF : Towards a mashup to build bioinformatics knowledge systems ». In : *Journal of Biomedical Informatics* 41.5, p. 706-716. doi : 10.1016/j.jbi.2008.03.004 (*citée page* 37).

Ben-Ahmed, W. (2005). « SAFE-NEXT : A SYSTEMIC APPROACH FOR KNOWLEDGE DISCOVERY IN DATABASES : APPLICATION IN ACCIDENT SCENARIO DEVELOPMENT AND INTERPRETATION ». Theses. Ecole Centrale Paris (*citée page* 67).

Berkovsky, S. et J. Freyne (2010). « Group-based recipe recommendations ». In : *Proceedings of the fourth ACM conference on Recommender systems - RecSys*. ACM Press. doi : 10.1145/1864708.1864732 (*citée page* 113).

Berners-Lee, T. (1994). « A Unifying Syntax for the Expression of Names and Addresses of Objects on the Network as used in the World-Wide Web ». In : *Universal Resource Identifiers in WWW* (*citée page* 23).

— (2006). *Linked Data*. W3C Working Draft. W3C (*citée page* 29).

Berners-Lee, T. et D. Connolly (2008). *Notation3 (N3) : A readable RDF syntax*. W3C Working Draft. W3C (*citée page* 27).

Berners-Lee, T. et D. Connolly (2003). *Hypertext Markup Language (HTML) : A Representation of Textual Information and MetaInformation for Retrieval and Interchange*. Internet-Draft draft-ietf-iiir-html-01. Work in Progress. Internet Engineering Task Force. 41 p. (*citée page* 45).

Bertin, J. (1999). *Sémiologie graphique les diagrammes, les réseaux, les cartes*. Les réimpressions. Paris : Editions de l'Ecole des Hautes Etudes en Sciences (31 janvier 1999), p. 431 (*citée page* 79, 80, 87).

Bertschi, S., S. Bresciani, T. Crawford, R. Goebel, W. Kienreich, M. Lindner, V. Sabol et A. V. Moere (2011). « What is Knowledge Visualization ? Perspectives on an Emerging Discipline ». In : *2011 15th International Conference on Information Visualisation*. IEEE. doi : 10.1109/iv.2011.58 (*citée page* 88).

Besnaci, M., N. Guin et P. Champin (2015). « Acquisition de connaissances pour importer des traces existantes dans un système de gestion de bases de traces ». In : *IC 2015 : 26es Journées francophones d'Ingénierie des Connaissances (Proceedings of the 26th French Knowledge Engineering Conference), Rennes, France, July 1-3, 2015.* (*citée page* 97).

Bianchini, D., V. D. Antonellis, N. D. Franceschi et M. Melchiori (2017). « PREFer : A prescription-based food recommender system ». In : *Computer Standards & Interfaces* 54, p. 64-75. doi : 10.1016/j.csi.2016.10.010 (*citée page* 113, 115, 118).

BioPortal (2018). *Welcome to BioPortal, the world's most comprehensive repository of biomedical ontologies. Dernière consultation : Septembre 2018*. url : https://bioportal.bioontology.org/ (*citée page* 31).

Bodenheimer, T. (2002). « Patient Self-management of Chronic Disease in Primary Care ». In : *JAMA* 288.19, p. 2469. doi : 10.1001/jama.288.19.2469 (*citée page* 120).

Bodenreider, O. (2006). « Bio-ontologies : current trends and future directions ». In : *Briefings in Bioinformatics* 7.3, p. 256-274. doi : 10.1093/bib/bbl027 (*citée page* 35).

Bognár, A. et J. Piekarski (2000). « Guidelines for Recipe Information and Calculation of Nutrient Composition of Prepared Foods (Dishes) ». In : *Journal of Food*

Composition and Analysis 13.4, p. 391-410. DOI : 10.1006/jfca.2000.0922 (*citée page* 137).

Bounhas, I. et Y. Slimani (2010). « A hierarchical approach for semi-structured document indexing and terminology extraction ». In : *2010 International Conference on Information Retrieval & Knowledge Management (CAMP)*, p. 315-320 (*citée page* 46).

Bray, T., J. Paoli et C. Sperberg-McQueen (1998). « Extensible Markup Language (XML) 1.0 : W3C Recommendation 10-February-1998 ». In : Springer Berlin Heidelberg, p. 22-32 (*citée page* 47).

Brooks, D. R. (2007). *An Introduction to HTML and JavaScript for Scientists and Engineers*. 1. Springer London, p. 200. DOI : 10.1007/978-1-84628-657-5 (*citée page* 45).

Brunzel, M. et M. Spiliopoulou (2006). « Discovering Multi Terms and Co-hyponymy from XHTML Documents with XTREEM ». In : *International Workshop on Knowledge Discovery from XML Documents : KDXD 2006 : Knowledge Discovery from XML Documents*. Springer Berlin Heidelberg, p. 22-32. DOI : 10.1007/11730262_5 (*citée page* 47).

Bundasak, S. et K. Chinnasarn (2013). « Dimensionality reduction on slope one predictor in the food recommender system ». In : *2013 International Computer Science and Engineering Conference (ICSEC)*. IEEE. DOI : 10.1109/icsec.2013.6694763 (*citée page* 118).

Burke, R. (2002). In : *User Modeling and User-Adapted Interaction* 12.4, p. 331-370. DOI : 10.1023/a:1021240730564 (*citée page* 111).

— (2007). « Hybrid Web Recommender Systems ». In : *The Adaptive Web*. Springer Berlin Heidelberg, p. 377-408. DOI : 10.1007/978-3-540-72079-9_12 (*citée page* 106, 111).

Burkhard, R. (2004). « Learning from architects :the difference between knowledge visualization and information visualization ». In : *Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004.* IEEE. DOI : 10.1109/iv.2004.1320194 (*citée page* 88, 89).

Burkhardt, D., K. Nazemi, J. D. Encarnacao, W. Retz et J. Kohlhammer (2014). « Visualization Adaptation Based on Environmental Influencing Factors ». In : *International Conference on Human-Computer Interaction HCI 2014 : Human-Computer Interaction. Theories, Methods, and Tools*. Springer International Publishing, p. 411-422. DOI : 10.1007/978-3-319-07233-3_38 (*citée page* 83).

Bussell, G. (2005). « Nutritional profiling vs guideline daily amounts as a means of helping consumers make appropriate food choices ». In : *Nutrition & Food Science* 35.5. Sous la dir. de M. Blades, p. 337-343. DOI : 10.1108/00346650510625548 (*citée page* 129).

Campos, L. M. de, J. M. Fernández-Luna et J. F. Huete (2008). « A collaborative recommender system based on probabilistic inference from fuzzy observations ». In : *Fuzzy Sets and Systems* 159.12, p. 1554-1576. DOI : 10.1016/j.fss.2008.01.016 (*citée page* 111).

Cantais, J., D. Dominguez, V. Gigante, L. Laera et V. Tamma (2005). « An Example of Food Ontology for Diabetes Control ». In : *Proceedings of the International Semantic Web Conference 2005 workshop on Ontology Patterns for the Semantic Web* (*citée page* 33).

Cao, N. et W. Cui (2016). *Introduction to Text Visualization*. 1. Atlantis Press, Paris, p. 114. DOI : 10.2991/978-94-6239-186-4 (*citée page* 83).

Capet, P., T. Delavallade, T. Nakamura, A. Sandor, C. Tarsitano et S. Voyatzi (2008). « A Risk Assessment System with Automatic Extraction of Event Types ». In : *IFIP – The International Federation for Information Processing*. Springer US, p. 220-229. doi : 10.1007/978-0-387-87685-6_27 (*citée page* 16).

Caracciolo, C., A. Stellato, A. Morshed, G. Johannsen, S. Rajbhandari, Y. Jaques et J. Keizer (2013). « The AGROVOC Linked Dataset ». In : *Semantic Web* 4.3, p. 341-348. doi : 10.3233/SW-130106 (*citée page* 32).

Card, S., J. Mackinlay et B. Shneiderman (1999). *Readings in Information Visualization : Using Vision to Think (Interactive Technologies)*. San Francisco, Calif : Morgan Kaufmann (*citée page* 82, 94).

Çelik, D. (2015). « FoodWiki : Ontology-Driven Mobile Safe Food Consumption System ». In : *The Scientific World Journal* 2015, p. 1-22. doi : 10.1155/2015/475410 (*citée page* 34).

Champin, P.-A., A. Mille et Y. Prié (2013). « Vers des traces numériques comme objets informatiques de premier niveau ». In : *Intellectica - La revue de l'Association pour la Recherche sur les sciences de la Cognition (ARCo)* 59, p. 171-204 (*citée page* 97).

Charlet, J. (2002). « Knowledge Engineering : developments, results and perspectives for Knowledge Management in medical field ». Habilitation à diriger des recherches. Université Pierre et Marie Curie - Paris VI (*citée page* 68).

Charlet, J., D. Bourigault, G. Kassel et M. Zacklad (2000). *Ingénierie des connaissances : évolutions récentes et nouveaux défis*. Collection technique et scientifique des télécommunications. Eyrolles (*citée page* 5, 67).

Charrondiere, U., B. Burlingame, S. Berman et I. Elmadfa (2011). *Food Composition Study Guide : Questions, Exercises and Answers*. Rome, Italy : FAO (*citée page* 136, 148).

Chen, C. h., W. Hardle et A. Unwin (2008). *Handbook of Data Visualization*. 1. Springer Berlin Heidelberg, p. 936. doi : 10.1007/978-3-540-33037-0 (*citée page* 83).

Chen, N., J. Suh, J. Verwey, G. Ramos, S. Drucker et P. Simard (2018). « AnchorViz : Facilitating Classifier Error Discovery Through Interactive Semantic Data Exploration ». In : *23rd International Conference on Intelligent User Interfaces*. Tokyo, Japan : ACM, p. 269-280. doi : 10.1145/3172944.3172950 (*citée page* 93).

Chen, Z., J. Hong, H. Li et Q. Xu (2008). « Web-Based Decision Support for Knowledge Sharing and Management in Intelligent Buildings Assessment ». In : *Encyclopedia of Decision Making and Decision Support Technologies*. IGI Global, p. 958-968. doi : 10.4018/978-1-59904-843-7.ch108 (*citée page* 88).

Chiang, J.-H., P.-C. Yang et H. Tu (2014). « Pattern analysis in daily physical activity data for personal health management ». In : *Pervasive and Mobile Computing* 13, p. 13-25. doi : 10.1016/j.pmcj.2013.12.003 (*citée page* 114).

CISMeF (2018). *Catalogue et Index des Sites Médicaux de langue Française. Dernière consultation : Septembre 2018*. url : http://www.chu-rouen.fr/cismef/ (*citée page* 5).

Clark, C. et S. Divvala (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*. (*citée page* 52).

Claveau, V. et A. Ncibi (2013). « Découverte de connaissances dans les séquences par CRF non-supervisés ». In : *20ème conférence sur le Traitement Automatique des Langues Naturelles, TALN*. T. 1. Sables d'Olonne, France, volume 1 (*citée page* 45).

CONROY, R., K. PYORALA, A. FITZGERALD, S. SANS, A. MENOTTI, G. De-Backer, D. De-Bacquer, P. DUCIMETIERE, P. JOUSILAHTI, U. KEIL, I. NJOLSTAD, R. OGANOV, T. THOMSEN, H. TUNSTALL-PEDOE, A. TVERDAL, H. WEDEL, P. WHINCUP, L. WILHELMSEN et I. GRAHAM (2003). « Estimation of ten-year risk of fatal cardiovascular disease in Europe : the SCORE project ». eng. In : *EUROPEAN HEART JOURNAL* 24.11, p. 987-1003 (*citée page* 18, 19).

Crestan, E. et P. Pantel (2010). « Web-scale knowledge extraction from semi-structured tables ». In : *Proceedings of the 19th international conference on World wide web - WWW '10*. ACM Press. doi : 10.1145/1772690.1772814 (*citée page* 46).

DAgostino, R., R. Vasan, M. Pencina, P. Wolf, M. Cobain, J. Massaro et W. Kannel (2008). « General Cardiovascular Risk Profile for Use in Primary Care : The Framingham Heart Study ». In : *Circulation* 117.6, p. 743-753. doi : 10.1161/circulationaha.107.699579 (*citée page* 18).

Damiani, E., C. Nocerino et M. Viviani (2004). « Knowledge extraction from an XML data flow : building a taxonomy based on clustering technique ». In : *EUROFUSE Workshop on Data and Knowledge Engineering (EUROFUSE 2004)* (*citée page* 47).

Darmon, N., J. Sondey, V. Azaïs-Braesco et M. Maillot (2017). « The SENS algorithm—a new nutrient profiling system for food labelling in Europe ». In : *European Journal of Clinical Nutrition* 72.2, p. 236-248. doi : 10.1038/s41430-017-0017-6 (*citée page* 130).

Darmoni, S., J. Leroy, F. Baudic, M. Douyère, J. Piot et B. Thirion (2000). « CISMeF : a structured health resource guide. » In : *Methods of information in medicine* 39.1, p. 30-35 (*citée page* 36).

Datko, P., W. D. Scherz, O. R. Velicu, R. Seepold et N. M. Madrid (2016). « Personal Recommendation System for Improving Sleep Quality ». In : *Intelligent Decision Technologies 2016*. Springer International Publishing, p. 405-412. doi : 10.1007/978-3-319-39630-9_34 (*citée page* 112).

Defanti, T., M. Brown et B. Mccormick (1989). « Visualization : expanding scientific and engineering research opportunities ». In : *Computer* 22.8, p. 12-16. doi : 10.1109/2.35195 (*citée page* 81).

Despres, S. (2014). « Construction d'une ontologie modulaire pour l'univers de la cuisine numérique ». In : *IC - 25èmes Journées francophones d'Ingénierie des Connaissances*. Sous la dir. de C. Faron-Zucker. Clermont-Ferrand, France, p. 27-38 (*citée page* 132, 134, 142).

Desrosiers, C. et G. Karypis (2010). « A Comprehensive Survey of Neighborhood based Recommendation Methods ». In : *Recommender Systems Handbook*. Springer US, p. 107-144. doi : 10.1007/978-0-387-85820-3_4 (*citée page* 110).

Diederich, J. et T. Iofciu (2006). « Finding Communities of Practice from User Profiles Based on Folksonomies. » In : *EC-TEL Workshops*. Sous la dir. d'E. Tomadaki et P. J. Scott. T. 213. CEUR Workshop Proceedings. CEUR-WS.org (*citée page* 108).

Ding, L., D. Difranzo, S. Magidson, D. Mcguinness et J. Hendler (2010). « Data-Gov Wiki : Towards Linked Government Data ». In : (*citée page* 46).

Dmitriev, D. (2018). « Digitizing National Cuisines : Cooking Recipes as Conceptual Graphs ». Anglais. In : *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Sous la dir. de C. Soria, L. Besacier et L. Pretorius. Miyazaki, Japan : European Language Resources Association (ELRA) (*citée page* 134).

El-Dosuky, M. A., M. Z. Rashad, T. T. Hamza et A. H. EL-Bassiouny (2012). « Food Recommendation Using Ontology and Heuristics ». In : p. 423-429. doi : 10.1007/978-3-642-35326-0_42 (*citée page* 118).

Ducharme, R. et P. Fraisse (1965). « Etude genetique de la memorisation de mots et d'images. » In : *Canadian Journal of Psychology/Revue canadienne de psychologie* 19.3, p. 253-261. doi : 10.1037/h0082907 (*citée page* 79).

Ducrot, P., C. Méjean, C. Julia, M. Touvier, E. Kesse-Guyot, L. Fezeu, S. Hercberg et S. Péneau (2014). « P224 : Compréhension objective vis-à-vis de différents systèmes d'information nutritionnelle simplifiés sur la face avant des emballages des aliments : étude NutriNet-Santé ». In : *Nutrition Clinique et Métabolisme* 28, S186. doi : 10.1016/s0985-0562(14)70866-0 (*citée page* 150).

Ducrot, P. (2016). « Approche épidémiologique des choix alimentaires lors de l'approvisionnement et de la préparation des repas : déterminants et associations avec l'alimentation et le tatut pondéral ». Thèse de doct. Université de Paris 13 (*citée page* 149).

Dumas, B., T. Broché, L. Hoste et B. Signer (2012). « ViDaX ». In : *Proceedings of the International Working Conference on Advanced Visual Interfaces - AVI '12*. ACM Press. doi : 10.1145/2254556.2254702 (*citée page* 94).

Dzemyda, G., O. Kurasova et J. Žilinskas (2013). *Multidimensional Data Visualization : Methods and Applications*. T. 75. 1. Springer New York, p. 252. doi : 10.1007/978-1-4419-0236-8 (*citée page* 83).

Eirinaki, M., M. Vazirgiannis et I. Varlamis (2003). « SEWeP ». In : *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03*. ACM Press. doi : 10.1145/956750.956765 (*citée page* 109).

Ekstrand, M. D., J. T. Riedl et J. A. Konstan (2011). « Collaborative Filtering Recommender Systems ». In : *Foundations and Trends in Human-Computer Interaction* 4.2, p. 81-173. doi : 10.1561/1100000009 (*citée page* 110).

Elmadfa, I. et A. L. Meyer (2010). « Importance of food composition data to nutrition and public health ». In : *European Journal of Clinical Nutrition* 64.S3, S4-S7. doi : 10.1038/ejcn.2010.202 (*citée page* 136).

Ge, M. et F. Ricci, éd. (2015). *Bringing the "healthy" into Food Recommenders*. T. 1533. CEUR Workshop Proceedings. CEUR-WS.org, p. 33-36 (*citée page* 113).

Elsweiler, D., B. Ludwig, A. Said, H. Schaefer et C. Trattner (2016). « Engendering Health with Recommender Systems ». In : *Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16*. ACM Press. doi : 10.1145/2959100.2959203 (*citée page* 112).

Eppler, M. J. et R. A. Burkhard (2004). « Knowledge Visualization Towards a New Discipline and its Fields of Application ». In : *ICA Working Paper 2/2004*. Institute for Corporate Communication (*citée page* 72).

Ermilov, I., S. Auer et C. Stadler (2013). « User-driven Semantic Mapping of Tabular Data ». In : *Proceedings of the 9th International Conference on Semantic Systems*. I-SEMANTICS '13. Graz, Austria : ACM, p. 105-112. doi : 10.1145/2506182.2506196 (*citée page* 52).

Ermine, J. L., M. Chaillot, P. Bigeon et B. Charreton (1996). « MKSM, a method for knowledge management ». In : 4.4, p. 541-575 (*citée page* 66, 68).

Erzsébet, M., P. Balazs, G. Balazs, V. Istvan, K. Gyorgy et N. Istvanné (2010). « A Formal Domain Model for Dietary and Physical Activity Counseling ». In : *Proceedings of the 14th International Conference on Knowledge-based and Intelligent Information and Engineering Systems : Part I*. KES'10. Cardiff, UK : Springer-Verlag, p. 607-616 (*citée page* 114).

Fadzli, S. A. et R. Setchi (2010). « Semantic Approach to Image Retrieval Using Statistical Models Based on a Lexical Ontology ». In : *Knowledge-Based and Intelligent Information and Engineering Systems*. Springer Berlin Heidelberg, p. 240-250. doi : 10.1007/978-3-642-15384-6_26 (*citée page* 144).

Farrell, R., C. Danis, S. Ramakrishnan et W. Kellogg (2012). « Intrapersonal retrospective recommendation : Lifestyle change recommendations using stable patterns of personal behavior ». In : 891, p. 24-28 (*citée page* 112).

Feinberg, M., J.-C. Favier et J. Ireland-Ripert (1992). « 15. Tables de composition des aliments : réalisation, utilisation, limites ». FRE. In : *Alimentation et nutrition humaines*. ESF, p. 1410-1438 (*citée page* 140).

Feinleib, M. (2001). « A Dictionary of Epidemiology, Fourth Edition - Edited by John M. Last, Robert A. Spasoff, and Susan S. Harris ». In : *American Journal of Epidemiology* 154.1, p. 93-94. doi : 10.1093/aje/154.1.93-a (*citée page* 15).

Feitosa, G. S. (2015). *Prevention of Cardiovascular Diseases*. Sous la dir. de J. Andrade, F. Pinto et D. Arnett. Springer International Publishing, p. 195-205. doi : 10.1007/978-3-319-22357-5 (*citée page* 19).

Fernandez-Luque, L., R. Karlsen et L. Vognild (2009). « Challenges and Opportunities of Using Recommender Systems for Personalized Health Education ». In : *Studies in Health Technology and Informatics* 150.Medical Informatics in a United and Healthy Europe, p. 903-907. doi : 10.3233/978-1-60750-044-5-903 (*citée page* 112, 113).

Ferrières, J. (2012). « Évaluation du risque cardiovasculaire ». In : *Archives of Cardiovascular Diseases Supplements* 4.4, p. 248-258. doi : 10.1016/s1878-6480(12)70840-6 (*citée page* 15).

Fierz, W. et R. Grütter (2000). « The SGML Standardization Framework and the Introduction of XML ». In : *Journal of Medical Internet Research* 2.2, e12. doi : 10.2196/jmir.2.2.e12 (*citée page* 47).

Finglas, P. M., R. Berry et S. Astley (2014). « Assessing and Improving the Quality of Food Composition Databases for Nutrition and Health Applications in Europe : The Contribution of EuroFIR ». In : *Advances in Nutrition* 5.5, 608S-614S. doi : 10.3945/an.113.005470 (*citée page* 135).

Fischer, F. et H. Mandl (2000). « Construction of Shared Knowledge in Face-to-Face and Computer-Mediated Cooperation ». In : *Paper presented at the AERA Annual Meeting, New Orleans, LA*, p. 18 (*citée page* 69).

Fluit, C., M. Sabou et F. van Harmelen (2006). « Ontology-Based Information Visualization : Toward Semantic Web Applications ». In : *Visualizing the Semantic Web*. Springer London, p. 45-58. doi : 10.1007/1-84628-290-x_3 (*citée page* 91).

FoodOn (2018). *FoodOn : A farm to fork food ontology. Dernière consultation : Septembre 2018*. url : https://foodontology.github.io/foodon/ (*citée page* 34).

Fortin, C. et R. Rousseau (1989). *Psychologie cognitive - une approche de traitement de l'information*. Télé-université (*citée page* 80).

Frati, F. et K.-L. Ma, éd. (2018). *Graph Drawing and Network Visualization*. 1. Springer International Publishing, p. 634. doi : 10.1007/978-3-319-73915-1 (*citée page* 83).

Freyne, J. et S. Berkovsky (2010). « Recommending Food : Reasoning on Recipes and Ingredients ». In : *User Modeling, Adaptation, and Personalization*. Springer Berlin Heidelberg, p. 381-386. doi : 10.1007/978-3-642-13470-8_36 (*citée page* 113, 118).

Fuchs, B. (2017). « Assister l'utilisateur à expliciter un modèle de trace avec l'analyse de concepts formels. » In : *28es Journées francophones d'Ingénierie des Connaissances IC 2017*. Sous la dir. de C. Roussey. Actes IC 2017 28es Journées francophones d'Ingénierie des Connaissances. Caen, France, p. 151-162 (*citée page* 97).

Fung, K. W. et O. Bodenreider (2012). « Knowledge Representation and Ontologies ». In : *Health Informatics*. Springer London, p. 255-275. doi : 10.1007/978-1-84882-448-5_14 (*citée page* 35).

Gandon, F., C. Faron-Zucker et O. Corby (2012). *Le Web sémantique : comment lier les données et les schémas sur le Web*. Paris : Dunod (*citée page* 24-27, 47).

Garcia-Penalvo, F. J., R. Colomo-Palacios, J. Garcia et R. Theron (2012). « Towards an ontology modeling tool. A validation in software engineering scenarios ». In : *Expert Systems with Applications* 39.13, p. 11468-11478. doi : 10.1016/j.eswa.2012.04.009 (*citée page* 92).

Gaye, B., C. Prugger, M. C. Perier, F. Thomas, M. Plichart, C. Guibout, C. Lemogne, B. Pannier, P. Boutouyrie, X. Jouven et J. P. Empana (2016). « High level of depressive symptoms as a barrier to reach an ideal cardiovascular health. The Paris Prospective Study III ». In : *Scientific Reports* 6.1. doi : 10.1038/srep18951 (*citée page* 11).

Ge, M., F. Ricci et D. Massimo (2015). « Health-aware Food Recommender System ». In : *RecSys* (*citée page* 113, 118).

Geleijnse, G., P. Nachtigall, P. van Kaam et L. Wijgergangs (2011). « A personalized recipe advice system to promote healthful choices ». In : *Proceedings of the 15th international conference on Intelligent user interfaces - IUI '11*. ACM Press. doi : 10.1145/1943403.1943487 (*citée page* 113, 118).

Georgeon, O., A. Mille et T. Bellet (2006). « Abstract : un outil et une méthodologie pour analyser une activité humaine médiée par un artefact technique complexe ». In : *Ingéniérie des Connaissances. Semaine de la connaissance*. Sous la dir. de R. Lehn, M. Harzallah, N. Aussenac-Gilles et J. Charlet. IC 2006 : 17es journées francophones d'Ingénierie des Connaissances : dans le cadre de la Semaine de la Connaissance 2006. Nantes, France (*citée page* 97).

Giampaoli, S. (2007). « CUORE : a sustainable cardiovascular disease prevention strategy ». In : *European Journal of Cardiovascular Prevention & Rehabilitation* 14.2, p. 161-162. doi : 10.1097/hjr.0b013e328157f3e5 (*citée page* 18).

Giampaoli, S., L. Palmieri, C. Donfrancesco, S. Panico, D. Vanuzzo, L. Pilotto, M. Ferrario, G. Cesana, A. Mattiello, D. Vanuzzo et et al. (2012). « Cardiovascular risk assessment in Italy : the CUORE Project risk score and risk chart ». In : *Italian Journal of Public Health* 4.2. doi : 10.2427/5885 (*citée page* 18).

Gilbert, J. K., M. Reiner et M. Nakhleh, éd. (2008). *Visualization : Theory and Practice in Science Education*. T. 3. 1. Springer Netherlands, p. 325. doi : 10.1007/978-1-4020-5267-5 (*citée page* 81).

Goldberg, K., T. Roeder, D. Gupta et C. Perkins (2001). « Eigentaste : A Constant Time Collaborative Filtering Algorithm ». In : *Information Retrieval* 4.2, p. 133-151. doi : 10.1023/a:1011419012209 (*citée page* 110).

Gotardo, R.-A., E. Rafael, H. Junior, S. D. Zorzo et P.-R.-M. Cereda (2013). « Approach to Cold-Start Problem in Recommender Systems in the Context of Web-Based Education ». In : *2013 12th International Conference on Machine Learning and Applications*. IEEE. doi : 10.1109/icmla.2013.199 (*citée page* 105).

Grabar, N. et P. Zweigenbaum (1999). « Acquisition automatique de connaissances morphologiques sur le vocabulaire médical ». In : *Actes de la 6ème conférence sur le Traitement Automatique des Langues Naturelles*. Cargèse, France : Association pour le Traitement Automatique des Langues (*citée page* 44).

GraphQL (2018). *GraphQL : A query language for your API. Dernière consultation : novembre 2018*. url : https://graphql.org/ (*citée page* 5).

Grätzer, G. (2016). « The PDF file format ». In : *More Math Into LaTeX*. Springer International Publishing, p. 299-305. doi : 10.1007/978-3-319-23796-1_11 (*citée page* 49).

Greenfield, H. et D. Southgate (2007). *Données sur la composition des aliments : production, gestion et utilisation*. Rome : Organisation des Nations Unies pour l'alimentation et l'agriculture, p. 308 (*citée page* 140).

Greenfield, H. et D. Southgate (2003). *Food Composition Data-Production, Management, and Use*. B.A. Burlingame and U.R. Charrondiere. FOOD et AGRICULTURE ORGANIZATION OF THE UNITED NATIONS, p. 295 (*citée page* 138).

Gruber, T. R. (1995). « Toward principles for the design of ontologies used for knowledge sharing ? » In : *International Journal of Human-Computer Studies* 43.5-6, p. 907-928. doi : 10.1006/ijhc.1995.1081 (*citée page* 25).

Hajifathalian, K., P. Ueda, Y. Lu, M. Woodward, A. Ahmadvand, C. A. Salinas, F. Azizi, R. Cifkova, M. D. Cesare, L. Eriksen, F. Farzadfar, N. Ikeda, D. Khalili, Y.-H. Khang, V. Lanska, L. León-Muñoz, D. Magliano, K. P. Msyamboza, K. Oh, F. Rodriguez-Artalejo, R. Rojas-Martinez, J. E. Shaw, G. A. Stevens, J. Tolstrup, B. Zhou, J. A. Salomon, M. Ezzati et G. Danaei (2015). « A novel risk score to predict cardiovascular disease risk in national populations (Globorisk) : a pooled analysis of prospective cohorts and health examination surveys ». In : *The Lancet Diabetes & Endocrinology* 3.5, p. 339-355. doi : 10.1016/s2213-8587(15)00081-9 (*citée page* 18).

Hamada, S. E. (2003). « Enrichment lexical knowledge with interword based features ». In : *IEEE International Conference on Systems, Man and Cybernetics*. IEEE. doi : 10.1109/icsmc.2002.1176105 (*citée page* 144).

Harish, D., Y. Srinivas, K. Rajesh et P. Anuradha (2011). « Image Annotations using Machine Learning and Features of ID3 Algorithm ». In : *International Journal of Computer Applications* 25.5, p. 45-49. doi : 10.5120/3024-4090 (*citée page* 143).

Harvey, M., B. Ludwig et D. Elsweiler (2012). « Learning user tastes : a first step to generating healthy meal plans ». In : *First international workshop on recommendation technologies for lifestyle change (lifestyle 2012)*, p. 18-23 (*citée page* 113).

Hawley, K. L., C. A. Roberto, M. A. Bragg, P. J. Liu, M. B. Schwartz et K. D. Brownell (2012). « The science on front-of-package food labels ». In : *Public Health Nutrition* 16.03, p. 430-439. doi : 10.1017/s1368980012000754 (*citée page* 130).

He, J., L. Liu, F. Yu et Y. Han (2016). « A Method of RDF Fuzzy Query Based on No Query Language Service with Permutated Breadth First Search Algorithm ». In :

Procedia Computer Science 100, p. 321-328. DOI : 10.1016/j.procs.2016.09.163 (*citée page* 144).

HEARST, M. A. (1999). *User interfaces and visualization*. Sous la dir. de B.-Y. R. A et R.-N. BERTHIER. Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc., p. 257-324 (*citée page* 83).

HEARST, M. (1992). « Automatic acquisition of hyponyms from large text corpora ». In : *Proceedings of the 14th conference on Computational linguistics*. Association for Computational Linguistics. DOI : 10.3115/992133.992154 (*citée page* 44).

HEIBERGER, R. M. et B. HOLLAND (2015). *Statistical Analysis and Data Display*. 2. Springer-Verlag New York, p. 898. DOI : 10.1007/978-1-4939-2122-5 (*citée page* 41).

HELMY, T., A. AL-NAZER, S. AL-BUKHITAN et A. IQBAL (2015). « Health, Food and User's Profile Ontologies for Personalized Information Retrieval ». In : *Procedia Computer Science* 52, p. 1071-1076. DOI : 10.1016/j.procs.2015.05.114 (*citée page* 113).

HENDLER, T. B.-L. J. et O. LASSILA (2001). « The Semantic Web ». In : *Scientific American*, p. 29-37 (*citée page* 21).

HENRY, K. et A. SENGUPTA (2007). « Extracting knowledge from XML document repository : a semantic Web-based approach ». In : *Information Technology and Management* 8.3, p. 205-221. DOI : 10.1007/s10799-007-0017-7 (*citée page* 47).

HEPP, M. (2008). « GoodRelations : An Ontology for Describing Products and Services Offers on the Web ». In : *Knowledge Engineering : Practice and Patterns*. Springer Berlin Heidelberg, p. 329-346. DOI : 10.1007/978-3-540-87696-0_29 (*citée page* 34).

HERCBERG, S. (2013). *Propositions pour un nouvel élan de la politique nutritionnelle française de santé publique*. Rapport 15/11/2013 pour Mme la ministre de la santé, dans le cadre de la Stratégie Nationale de Santé, France. France (*citée page* 128-130, 138, 149).

HERLOCKER, J. L., J. A. KONSTAN, L. G. TERVEEN et J. T. RIEDL (2004). « Evaluating collaborative filtering recommender systems ». In : *ACM Transactions on Information Systems* 22.1, p. 5-53. DOI : 10.1145/963770.963772 (*citée page* 116).

HERMAN, I., G. MELANCON et M. MARSHALL (2000). « Graph visualization and navigation in information visualization : A survey ». In : *IEEE Transactions on Visualization and Computer Graphics* 6.1, p. 24-43. DOI : 10.1109/2945.841119 (*citée page* 89).

HIEMSTRA, D., N. KUSHMERICK, C. DOMENICONI, C. D. PAICE, M. W. CARROLL, C. S. JENSEN, F. GIUNCHIGLIA, I. ZAIHRAYEU, D. ZHANG, Y. MANOLOPOULOS, Y. THEODORIDIS, V. J. TSOTRAS, J. SHEN, M. O. WARD, A. DATTA, N. TATBUL, P. MITRA, P. BARCELÓ, S. A. BRIDWELL, H. J. MILLER, R. SCHENKEL, E. RAHM, A. REUTER, A. BORGIDA, M. A. CASANOVA, A. H. F. LAENDER, C. KOCH, F. CRESTANI, T. WESTERVELD, K. VORUGANTI, K. VORUGANTI, K. WADA, J. SHEN, N. PALMER et S. MANKOVSKII (2009). « Linking and Brushing ». In : *Encyclopedia of Database Systems*. Springer US, p. 1623-1626. DOI : 10.1007/978-0-387-39940-9_1129 (*citée page* 83).

HIPPISLEY-COX, J., C. COUPLAND, Y. VINOGRADOVA, J. ROBSON, M. MAY et P. BRINDLE (2007). « Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom : prospective open cohort study ». In : *BMJ* 335.7611, p. 136. DOI : 10.1136/bmj.39261.471806.55 (*citée page* 18).

Hong, C. (2003). « Compound Brushing ». In : *Proceedings of the Ninth Annual IEEE Conference on Information Visualization*. INFOVIS'03. Seattle, Washington : IEEE Computer Society, p. 181-188 (*citée page* 83).

Hors-Fraile, S., F. Benjumea, L.-C. Hernandez, F. O. Ruiz et L. Fernandez (2016). « Design of two combined health recommender systems for tailoring messages in a smoking cessation app ». In : *CoRR* abs/1608.07192 (*citée page* 117).

Hussain, A., K. Latif, A. T. Rextin, A. Hayat et M. Alam (2014). « Scalable Visualization of Semantic Nets using Power-Law Graphs ». In : *Applied Mathematics & Information Sciences* 8.1, p. 355-367. doi : 10.12785/amis/080145 (*citée page* 91, 92).

Hussain, A., K. Farooq, B. Luo et W. Slack (2015). « A Novel Ontology and Machine Learning Inspired Hybrid Cardiovascular Decision Support Framework ». In : *2015 IEEE Symposium Series on Computational Intelligence*. IEEE. doi : 10.1109/ssci.2015.122 (*citée page* 18).

HyperGraphQL (2018). *HyperGraphQL : A GraphQL interface for querying and serving linked data on the Web. Dernière consultation : novembre 2018*. url : https://www.hypergraphql.org/ (*citée page* 5).

IEEE, éd. (1991). *610-1991*. doi : 10.1109/ieeestd.1991.106963 (*citée page* 22).

Ireland, J. et A. Møller (2013). « What's New in LanguaL™ ? » In : *Procedia Food Science* 2, p. 117-121. doi : 10.1016/j.profoo.2013.04.018 (*citée page* 135, 153).

Iversen, G. R. et M. Gergen (1997). *Statistics : The Conceptual Approach*. 1. Springer-Verlag New York, p. 735. doi : 10.1007/978-1-4612-2244-6 (*citée page* 40).

Izumi, S., D. Kuriyama, G. Itabashi, A. Togashi, Y. Kato et K. Takahashi (2006). « An ontology-based advice system for health and exercise ». In : *Proceedings of the Tenth IASTED International Conference on Internet and Multimedia Systems and Applications (IMSA 2006), Honolulu, Hawaii, USA, August 14-16, 2006*, p. 95-100 (*citée page* 114).

Jian, W., J. Killian, H. Yang, K. Williams, S. R. Choudhury, S. Tuarob, C. Caragea et C. Giles (2015). « PDFMEF ». In : *Proceedings of the Knowledge Capture Conference on ZZZ-K-CAP 2015*. ACM Press. doi : 10.1145/2815833.2815834 (*citée page* 49).

Jilani, I. et F. Amardeilh (2009). « Enrichissement automatique d'une base de connaissances biologiques à l'aide des outils du Web sémantique ». In : *IC 2009 - 20èmes journées francophones d'Ingénierie des Connaissances*. Hammamet, Tunisia, p. 47 (*citée page* 44).

Jonassen, D. et R. Marra (2011). « Concept mapping and other formalisms as Mindtools for representing knowledge ». In : *Research in Learning Technology* 2.1. doi : 10.3402/rlt.v2i1.9573 (*citée page* 69).

Jonquet, C., A. Toulet, E. Arnaud, S. Aubin, E. D. Yeumo, V. Emonet, J. Graybeal, M.-A. Laporte, M. A. Musen, V. Pesce et P. Larmande (2018). « AgroPortal : A vocabulary and ontology repository for agronomy ». In : *Computers and Electronics in Agriculture* 144, p. 126-143. doi : 10.1016/j.compag.2017.10.012 (*citée page* 31).

Julia, C., P. Ducrot, S. Péneau, V. Deschamps, C. Méjean, L. Fézeu, M. Touvier, S. Hercberg et E. Kesse-Guyot (2015). « Discriminating nutritional quality of foods using the 5-Color nutrition label in the French food market : consistency with nutritional recommendations ». In : *Nutrition Journal* 14.1. doi : 10.1186/s12937-015-0090-4 (*citée page* 130).

JUNYONG, I. et L. SANGSEOK (2017). « Statistical data presentation ». In : *Korean Journal of Anesthesiology* 70.3, p. 267. DOI : 10.4097/kjae.2017.70.3.267 (*citée page* 41).

KAO, M.-Y., éd. (2016). *Encyclopedia of Algorithms*. 2. Springer-Verlag New York, p. 2389. DOI : 10.1007/978-1-4939-2864-4 (*citée page* 86).

KARATZOGLOU, A., X. AMATRIAIN, L. BALTRUNAS et N. OLIVER (2010). « Multiverse Recommendation : N-dimensional Tensor Factorization for Context-aware Collaborative Filtering ». In : *Proceedings of the Fourth ACM Conference on Recommender Systems*. RecSys '10. Barcelona, Spain : ACM, p. 79-86. DOI : 10.1145/1864708.1864727 (*citée page* 111).

KATZ, D. L., V. Y. NJIKE, Z. FARIDI, L. Q. RHEE, R. S. REEVES, D. J. A. JENKINS et K. T. AYOOB (2009). « The Stratification of Foods on the Basis of Overall Nutritional Quality : The Overall Nutritional Quality Index ». In : *American Journal of Health Promotion* 24.2, p. 133-143. DOI : 10.4278/ajhp.080930-quan-224 (*citée page* 130).

KERREN, A., H. PURCHASE et M. WARD, éd. (2014). *Multivariate Network Visualization*. 1. Springer International Publishing, p. 237. DOI : 10.1007/978-3-319-06793-3 (*citée page* 83).

KERZNER, E., A. LEX, C. SIGULINSKY, T. URNESS, B. JONES, R. MARC et M. MEYER (2017). « Graffinity : Visualizing Connectivity in Large Graphs ». In : *Computer Graphics Forum* 36.3, p. 251-260. DOI : 10.1111/cgf.13184 (*citée page* 87).

KIM, B. M. et Q. LI (2004). « Probabilistic Model Estimation for Collaborative Filtering Based on Items Attributes ». In : *IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)*. T. 1. IEEE, p. 185-191. DOI : 10.1109/wi.2004.10066 (*citée page* 111).

KIM, J. K. et S. KANG (2017). « Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis ». In : *Journal of Healthcare Engineering* 2017, p. 1-13. DOI : 10.1155/2017/2780501 (*citée page* 19).

KITAMURA, K., C. de SILVA, T. YAMASAKI et K. AIZAWA (2010). « Image processing based approach to food balance analysis for personal food logging ». In : *2010 IEEE International Conference on Multimedia and Expo*. IEEE, p. 625-630. DOI : 10.1109/icme.2010.5583021 (*citée page* 137).

KLAMPFL, S., K. JACK et R. KERN (2014). « A Comparison of Two Unsupervised Table Recognition Methods from Digital Scientific Articles ». In : *D-Lib Magazine* 20.11/12. DOI : 10.1045/november14-klampfl (*citée page* 49).

KLYNE, G. et J. CARROL (2004). *Resource Description Framework (RDF) : Concepts and Abstract Syntax*. W3C Working Draft. W3C (*citée page* 5, 26).

KOLCHIN, M. et D. ZAMULA (2013). « Food Product Ontology : Initial Implementation of a Vocabulary for Describing Food Products ». In : *In Proceeding of the 14th Conference of Open Innovations Association FRUCT* 11-15, p. 191-196 (*citée page* 34).

KONSTAN, J. A. et J. RIEDL (2012). « Recommender systems : from algorithms to user experience ». In : *User Modeling and User-Adapted Interaction* 22.1-2, p. 101-123. DOI : 10.1007/s11257-011-9112-x (*citée page* 104).

KOREN, Y. (2008). « Factorization Meets the Neighborhood : A Multifaceted Collaborative Filtering Model ». In : *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '08. Las Vegas, Nevada, USA : ACM, p. 426-434. DOI : 10.1145/1401890.1401944 (*citée page* 110).

Koren, Y. et R. Bell (2010). « Advances in Collaborative Filtering ». In : *Recommender Systems Handbook*. Springer US, p. 145-186. doi : 10.1007/978-0-387-85820-3_5 (*citée page* 110).

— (2015). « Advances in Collaborative Filtering ». In : *Recommender Systems Handbook*. Springer US, p. 77-118. doi : 10.1007/978-1-4899-7637-6_3 (*citée page* 110).

Krones, T., H. Keller, A. Sonnichsen, E.-M. Sadowski, E. Baum, K. Wegscheider, J. Rochon et N. Donner-Banzhoff (2008). « Absolute Cardiovascular Disease Risk and Shared Decision Making in Primary Care : A Randomized Controlled Trial ». In : *The Annals of Family Medicine* 6.3, p. 218-227. doi : 10.1370/afm.854 (*citée page* 18).

Kung, H.-Y., T. M. P. Nguyen, T.-H. Kuo, C.-P. Tsai et C.-H. Chen (2011). « Intelligent Personalized Food Recommendation System Based on a Semantic Sensor Web ». In : *2011 International Conference in Electrics, Communication and Automatic Control Proceedings*. Springer New York, p. 61-68. doi : 10.1007/978-1-4419-8849-2_9 (*citée page* 114).

Lamy, J.-B. (2017). « Représentation, iconisation et visualisation des connaissances : Principes et applications à l'aide à la décision médicale ». Habilitation à diriger des recherches. Normandie Université, Université de Rouen (*citée page* 88).

Lee, W., C. Chen, J. Huang et J. Liang (2017). « A smartphone based activity aware system for music streaming recommendation ». In : *Knowledge Based Systems* 131, p. 70-82. doi : 10.1016/j.knosys.2017.06.002 (*citée page* 104).

Lehmann, U., V. R. Charles, A. Vlassopoulos, G. Masset et J. Spieldenner (2017). « Nutrient profiling for product reformulation : public health impact and benefits for the consumer ». In : *Proceedings of the Nutrition Society* 76.03, p. 255-264. doi : 10.1017/s0029665117000301 (*citée page* 128).

Li, H., G. Grinstein et L. Costello (2009). « A visual canonical adjacency matrix for graphs ». In : *2009 IEEE Pacific Visualization Symposium*. IEEE. doi : 10.1109/pacificvis.2009.4906842 (*citée page* 87).

Lieberman, H. (1995). « Letizia : An Agent That Assists Web Browsing ». In : *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*. IJCAI'95. Montreal, Quebec, Canada : Morgan Kaufmann Publishers Inc., p. 924-929 (*citée page* 108).

Liebig, T. et O. Noppens (2005). « OntoTrack : A semantic approach for ontology authoring ». In : *Web Semantics : Science, Services and Agents on the World Wide Web* 3.2-3, p. 116-131. doi : 10.1016/j.websem.2005.06.004 (*citée page* 90, 91).

Lin, C.-J., T.-T. Kuo et S.-D. Lin (2014). « A Content-Based Matrix Factorization Model for Recipe Recommendation ». In : *Advances in Knowledge Discovery and Data Mining*. Springer International Publishing, p. 560-571. doi : 10.1007/978-3-319-06605-9_46 (*citée page* 118).

Linden, G., B. Smith et J. York (2003). « Amazon.com recommendations : item-to-item collaborative filtering ». In : *IEEE Internet Computing* 7.1, p. 76-80. doi : 10.1109/mic.2003.1167344 (*citée page* 110).

Linn, M. (2003). « Technology and science education : Starting points, research programs, and trends ». In : *International Journal of Science Education* 25.6, p. 727-758. doi : 10.1080/09500690305017 (*citée page* 81).

Linsen, L. et S. Behrendt (2011). « Linked treemap : a 3D treemap-nodelink layout for visualizing hierarchical structures ». In : *Computational Statistics* 26.4, p. 679-697. doi : 10.1007/s00180-011-0272-2 (*citée page* 86).

Liu, D., K. Chen, Y. Chou et J. Lee (2018). « Online Recommendations based on Dynamic Adjustment of Recommendation Lists ». In : Elsevier BV. doi : 10.1016/j.knosys.2018.07.038 (*citée page* 104).

Liu, D., C. Lai et Y. Chen (2012). « Document recommendations based on knowledge flows : A hybrid of personalized and group-based approaches ». In : *Journal of the American Society for Information Science and Technology* 63.10, p. 2100-2117. doi : 10.1002/asi.22705 (*citée page* 104).

Lohmann, S., S. Negru, F. Haag et T. Ertl (2016). « Visualizing ontologies with VOWL ». In : *Semantic Web* 7.4. Sous la dir. de K. Janowicz et S. SchlobachEditors, p. 399-419. doi : 10.3233/SW-150200 (*citée page* 91-93).

Long, L. K., L. C. Hui, G. Y. Fook et W. M. N. W. Zainon (2017). « A Study on the Effectiveness of Tree-Maps as Tree Visualization Techniques ». In : *Procedia Computer Science* 124, p. 108-115. doi : 10.1016/j.procs.2017.12.136 (*citée page* 83).

Lops, P., M. de Gemmis et G. Semeraro (2010). « Content-based Recommender Systems : State of the Art and Trends ». In : *Recommender Systems Handbook*. Springer US, p. 73-105. doi : 10.1007/978-0-387-85820-3_3 (*citée page* 106, 108, 109).

LOV (2018). *Linked Open Vocabularies (LOV). Dernière consultation : Septembre 2018*. url : https://lov.linkeddata.es/dataset/lov (*citée page* 31).

Lovric, M. (2011). *International Encyclopedia of Statistical Science*. T. 2. Springer Berlin Heidelberg, p. 1671. doi : 10.1007/978-3-642-04898-2 (*citée page* 40).

Machorro-Cano, I., U. Ramos-Deonati, G. Alor-Hernández, J. L. Sánchez-Cervantes, C. Sánchez-Ramirez, L. Rodriguez-Mazahua et M. G. Segura-Ozuna (2017). « An IoT-Based Architecture to Develop a Healthcare Smart Platform ». In : *Communications in Computer and Information Science*. Springer International Publishing, p. 133-145. doi : 10.1007/978-3-319-67283-0_10 (*citée page* 17).

Madeleine, M. et V. Pietri (2000). *L'ingénierie de la connaissance : la nouvelle épistémologie appliquée*. Annales littéraires de l'Universite de Franche-Comté : Université de Franche-Comté vol.696. Presses universitaires franc-comtoises (*citée page* 66).

Magesh (2013). « COMPARING THE PERFORMANCE OF SEMANTIC IMAGE RETRIEVAL USING SPARQL QUERY, DECISION TREE ALGORITHM AND LIRE ». In : *Journal of Computer Science* 9.8, p. 1041-1050. doi : 10.3844/jcssp.2013.1041.1050 (*citée page* 143).

Magnini, B. et C. Strapparava, éd. (2001). *Improving User Modelling with Content-Based Techniques*. UM 01 : Proceedings of the 8th International Conference on User Modeling 2001. Berlin, Heidelberg : Springer-Verlag (*citée page* 109).

Mahesh, A., C. Pramila et C. Prithviraj (2012). « Information Extraction from Web Tables ». In : 2, p. 313-318 (*citée page* 46).

Malhotra, A., E. Younesi, M. Gündel, B. Müller, M. T. Heneka et M. H. Apitius (2014). « ADO : A disease ontology representing the domain knowledge specific to Alzheimer's disease ». In : *Alzheimer's & Dementia* 10.2, p. 238-246. doi : 10.1016/j.jalz.2013.02.009 (*citée page* 35).

Malintha, F., V. S. A. Cooray, G. H. T. Indeewara, K. S. Jayasinghe et F. Subha (2015). « Semi-supervised learning framework for knowledge extraction in Cricket domain ». In : *ITRU Research Symposium* (*citée page* 45).

Mallat, S., A. Zouaghi, E. Hkiri et M. Zrigui (2013). « Method of Lexical Enrichment in Information Retrieval System in Arabic ». In : *International Journal of*

Information Retrieval Research 3.4, p. 35-51. DOI : 10.4018/ijirr.2013100103 (*citée page* 144).

MARCONI, S., A. DURAZZO, E. CAMILLI, S. LISCIANI, P. GABRIELLI, A. AGUZZI, L. GAMBELLI, M. LUCARINI et L. MARLETTA (2018). « Food Composition Databases : Considerations about Complex Food Matrices ». In : *Foods* 7.1, p. 2. DOI : 10.3390/foods7010002 (*citée page* 134).

MARCOPULOS, B. (2017). « Recommendation ». In : *Encyclopedia of Clinical Neuropsychology*. Sous la dir. de J. KREUTZER, J. DELUCA et B. CAPLAN. Cham : Springer International Publishing, p. 1-1. DOI : 10.1007/978-3-319-56782-2_422-2 (*citée page* 105).

MARSHALL, S. et E. PRUD'HOMMEAUX (2008). *A Prototype Knowledge Base for the Life Sciences : W3C Interest Group Note 4 June 2008*. W3C Working Draft. W3C (*citée page* 37).

MASERAT, E., R. SAFDARI, H. AGHDAEI, A. KHOSHSIRAT et M. ZALI (2017). « 43 : DESIGNING EVIDENCE BASED RISK ASSESSMENT SYSTEM FOR CANCER SCREENING AS AN APPLICABLE APPROACH FOR THE ESTIMATING OF TREATMENT ROADMAP ». In : *BMJ Open* 7.Suppl 1. DOI : 10.1136/bmjopen-2016-015415.43 (*citée page* 17).

MAXWELL, A. (1978). *Basic Statistics For medical and social science students*. Springer Netherlands, p. 126. DOI : 10.1007/978-94-009-5804-3 (*citée page* 40).

MAZOUNI, M. (2008). « Towards a better approach of the risk management : from the ontological modeling of the accidental process to decision aided interactive system. » Theses. Institut National Polytechnique de Lorraine - INPL (*citée page* 16).

MCGRAW, T. (2015). « Graph-Based Visualization of Neuronal Connectivity Using Matrix Block Partitioning and Edge Bundling ». In : *Advances in Visual Computing*. Springer International Publishing, p. 3-13. DOI : 10.1007/978-3-319-27857-5_1 (*citée page* 87).

MCGUINNESS, D. L. et F. VAN-HARMELEN (2004). *OWL Web Ontology Language Overview*. W3C Working Draft. W3C (*citée page* 27, 28).

MCGUINNESS, D.-L. et F. VAN-HARMELEN (2012). *OWL 2 Web Ontology Language Document Overview (Second Edition)*. W3C Working Draft. W3C (*citée page* 29).

MEDDRA (2018). *MedDRA Medical Dictionary for Regulatory Activities. Dernière consultation : Septembre 2018*. URL : https://www.meddra.org/ (*citée page* 35).

MEILENDER, T. (2013). « A Semantic Wiki for Decision Knowledge Management : Application in Oncology ». Thèse de doct. Université de Lorraine (*citée page* 25, 33).

MENETON, P., C. LEMOGNE, E. HERQUELOT, S. BONENFANT, M. LARSON, R. VASAN, J. MÉNARD, M. GOLDBERG et M. ZINS (2016). « A Global View of the Relationships between the Main Behavioural and Clinical Cardiovascular Risk Factors in the GAZEL Prospective Cohort ». In : *PLOS ONE* 11.9, p. 1-20. DOI : 10.1371/journal.pone.0162386 (*citée page* 56, 73, 98).

MERZ, T. (1998). *Web Publishing with Acrobat/PDF*. Springer Berlin Heidelberg, p. 236. DOI : 10.1007/978-3-642-72032-1 (*citée page* 49).

MILES, A., B. MATTHEWS, M. WILSON et D. BRICKLEY (2005). « SKOS Core : Simple Knowledge Organisation for the Web ». In : *Proceedings of the 2005 International Conference on Dublin Core and Metadata Applications : Vocabularies in Practice*. DCMI '05. Madrid, Spain : Dublin Core Metadata Initiative, p. 1-9 (*citée page* 32).

MILLER, S., J. GUINNESS et A. ZAMANIAN (2004). « Name tagging with word clusters and discriminative training ». In : *Proceedings of HLT*, p. 337-342 (*citée page* 145).

Min, H., F. J. Manion, E. Goralczyk, Y.-N. Wong, E. Ross et J. R. Beck (2009). « Integration of prostate cancer clinical data using an ontology ». In : *Journal of Biomedical Informatics* 42.6, p. 1035-1045. doi : 10.1016/j.jbi.2009.05.007 (*citée page* 35).

Miranda, C. et A. M. Jorge (2009). « Item-Based and User-Based Incremental Collaborative Filtering for Web Recommendations ». In : *Progress in Artificial Intelligence*. Springer Berlin Heidelberg, p. 673-684. doi : 10.1007/978-3-642-04686-5_55 (*citée page* 110).

Møller, A. et J. Ireland (2018). *LanguaL 2017-Multilingual Thesaurus*. en. Technical Report. doi : 10.13140/rg.2.2.13274.64964 (*citée page* 32, 153).

MÖLLER, N., S. O. HANSSON et M. PETERSON (2006). « Safety is more than the antonym of risk ». In : *Journal of Applied Philosophy* 23.4, p. 419-432. doi : 10.1111/j.1468-5930.2006.00345.x (*citée page* 14).

Mooney, R. J. et L. Roy (2000). « Content-based Book Recommending Using Learning for Text Categorization ». In : *Proceedings of the Fifth ACM Conference on Digital Libraries*. DL '00. San Antonio, Texas, USA : ACM, p. 195-204. doi : 10.1145/336597.336662 (*citée page* 104).

Motik, B., P. Patel-Schneider et B. Parsia (2012). *OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition)*. W3C Working Draft. W3C (*citée page* 5, 29).

Mulrooney, H. M. et J. Bell (2016). « Does the food retail environment reflect UK public health recommendations for healthy eating ? » In : *Public Health* 134, p. 114-116. doi : 10.1016/j.puhe.2015.12.004 (*citée page* 113).

Muñoz, E., A. Hogan et A. Mileo (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*. ACM Press. doi : 10.1145/2556195.2556266 (*citée page* 46).

Munzner, T. (2009). « A Nested Model for Visualization Design and Validation ». In : *IEEE Transactions on Visualization and Computer Graphics* 15.6, p. 921-928. doi : 10.1109/tvcg.2009.111 (*citée page* 90).

Nakai, A. et K. Suzuki (2014). « Risk Assessment System for Verifying the Safeguards Based on the HAZOP Analysis ». In : *International Journal of Advanced Computer Science and Applications* 5.10. doi : 10.14569/ijacsa.2014.051008 (*citée page* 16).

Nakamura-Delloye, Y. et R. Stern (2011). « Extraction de relations et de patrons de relations entre entités nommées en vue de l'enrichissement d'une ontologie ». In : *TOTh 2011 : Terminologie & Ontologie : Théories et Applications*. Annecy, France, p. 50 (*citée page* 45).

Nakamura-Delloye, Y. et D.-L.-C. Villemonte (2010). « Exploitation de résultats d'analyse syntaxique pour extraction semi-supervisée des chemins de relations ». In : *17e Conférence sur le Traitement Automatique des Langues Naturelles-TALN 2010*. Montréal, Canada (*citée page* 45).

Nakamura, A. et N. Abe (1998). « Collaborative Filtering Using Weighted Majority Prediction Algorithms ». In : *Proceedings of the Fifteenth International Conference on Machine Learning*. ICML 98. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., p. 395-403 (*citée page* 110).

Namici, M. (2018). « R2RML Mappings in OBDA Systems : Enabling Comparison among OBDA Tools ». In : *CoRR* abs/1804.01405 (*citée page* 48).

NARAIN, R., S. SAXENA et A. GOYAL (2016). « Cardiovascular risk prediction : a comparative study of Framingham and quantum neural network based approach ». In : *Patient Preference and Adherence* Volume 10, p. 1259-1270. DOI : 10.2147/ppa.s108203 (*citée page* 19).

NAYAK, R. et M. J-ZAKI, éd. (2006). *Knowledge Discovery from XML Documents*. Springer Berlin Heidelberg. DOI : 10.1007/11730262 (*citée page* 47).

NAZEMI, K., D. BURKHARDT, E. GINTERS et J. KOHLHAMMER (2015). « Semantics Visualization – Definition, Approaches and Challenges ». In : *Procedia Computer Science* 75, p. 75-83. DOI : 10.1016/j.procs.2015.12.216 (*citée page* 89).

NAZER, A. A., T. HELMY et M. A. MULHEM (2014). « Users Profile Ontology-based Semantic Framework for Personalized Food and Nutrition Recommendation ». In : *Procedia Computer Science* 32, p. 101-108. DOI : 10.1016/j.procs.2014.05.403 (*citée page* 113, 118).

NCBI (2018). *PubMed. Dernière consultation : Septembre 2018*. URL : https://www.ncbi.nlm.nih.gov/pubmed/ (*citée page* 5, 33).

NELSON, T. (1965). « Complex information processing : a file structure for the complex, the changing and the indeterminate ». In : *ACM 65 Proceedings of the 1965 20th national conference*. ACM Press, p. 84-100. DOI : 10.1145/800197.806036 (*citée page* 22).

NETISOPAKUL, P., S. NAPAT, M. BURANARACH, Y. THEIN et T. SUPNITHI (2010). « A Knowledge-based Framework for Development of Personalized Food Recommender System ». In : *The Fifth International Conference on Knowledge, Information and Creativity Support Systems Novermber 25-27, 2010, Chiang Mai, Thailand* (*citée page* 116, 118).

NEURATH, O. (1936). *International picture language : the first rules of Isotype*. Sous la dir. de K. P. T. T. &. C. L. 1936. Psyche miniatures, General series, no. 83, p. 117 (*citée page* 79).

NEWELL, A. (1982). « The knowledge level ». In : *Artificial Intelligence* 18.1, p. 87-127. DOI : 10.1016/0004-3702(82)90012-1 (*citée page* 68).

NIJMAN, C. A. J., I. M. ZIJP, A. SIERKSMA, A. J. C. ROODENBURG, R. LEENEN, C. van den KERKHOFF, J. A. WESTSTRATE et G. W. MEIJER (2006). « A method to improve the nutritional quality of foods and beverages based on dietary recommendations ». In : *European Journal of Clinical Nutrition* 61.4, p. 461-471. DOI : 10.1038/sj.ejcn.1602548 (*citée page* 130).

NIKULIN, V. (2014). « Hybrid Recommender System for Prediction of the Yelp Users Preferences ». In : *Advances in Data Mining. Applications and Theoretical Aspects*. Springer International Publishing, p. 85-99. DOI : 10.1007/978-3-319-08976-8_7 (*citée page* 118).

NOUVEL, D., J.-Y. ANTOINE, N. FRIBURGER et A. SOULET (2013). « Fouille de règles d'annotation pour la reconnaissance d'entités nommées ». In : *Traitement Automatique des Langues* 54.2, p. 13-41 (*citée page* 144).

NOVAK, J. D., B. D. GOWIN et J. B. KAHLE (1984). *Learning how to learn*. Cambridge University Press, p. 55-76. DOI : 10.1017/cbo9781139173469 (*citée page* 69).

NOY, N., R. FERGERSON et M. MUSEN (2000). « The Knowledge Model of Protégé-2000 : Combining Interoperability and Flexibility ». In : *Knowledge Engineering and Knowledge Management Methods, Models, and Tools*. Springer Berlin Heidelberg, p. 17-32. DOI : 10.1007/3-540-39967-4_2 (*citée page* 90).

NUGUES, P. (2016). « An Overview of Knowledge Extraction Projects in the NLP group at Lund University ». eng. In : *Digital Humanities 2016. From Digitization*

to Knowledge 2016 : Resources and Methods for Semantic Processing of Digital Works/Texts, Proceedings of the Workshop, p. 25-31 (*citée page* 48).

O'Donovan, J. et B. Smyth (2005). « Trust in Recommender Systems ». In : *Proceedings of the 10th International Conference on Intelligent User Interfaces*. IUI '05. San Diego, California, USA : ACM, p. 167-174. doi : 10.1145/1040830.1040870 (*citée page* 116, 117).

Oh, Y., A. Choi et W. Woo (2009). « u-BabSang : a context-aware food recommendation system ». In : *The Journal of Supercomputing* 54.1, p. 61-81. doi : 10.1007/s11227-009-0314-5 (*citée page* 118).

Open-Food-Facts (2018). *Open Food Facts : l'information alimentaire ouverte. Dernière consultation : Septembre 2018*. url : https://fr.openfoodfacts.org/ (*citée page* 35).

Paananen, J. et G. Wong (2009). « FORG3D : Force-directed 3D graph editor for visualization of integrated genome scale data ». In : *BMC Systems Biology* 3.1, p. 26. doi : 10.1186/1752-0509-3-26 (*citée page* 86).

Paivio, A. (1969). « Mental imagery in associative learning and memory. » In : *Psychological Review* 76.3, p. 241-263. doi : 10.1037/h0027272 (*citée page* 79).

— (1974). « Language and Knowledge of the World ». In : *Educational Researcher* 3.9, p. 5-12. doi : 10.3102/0013189x003009005 (*citée page* 81).

Pan, C. et W. Li (2010). « Research paper recommendation with topic analysis ». In : *2010 International Conference On Computer Design and Applications*. IEEE, p. 264-268. doi : 10.1109/iccda.2010.5541170 (*citée page* 104).

Panchenko, A., P. Romanov, O. Morozova, H. Naets, A. Philippovich, A. Romanov et C. Fairon (2013). « Serelex : Search and Visualization of Semantically Related Words ». In : *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, p. 837-840. doi : 10.1007/978-3-642-36973-5_97 (*citée page* 93).

Paquette, G. (2002). *Modélisation des connaissances et des compétences : un langage graphique pour concevoir et apprendre*. Sainte-Foy Que : Presses de l'Université du Québec (*citée page* 67, 68).

Parra, D., P. Brusilovsky et C. Trattner (2014). « See What You Want to See : Visual User-driven Approach for Hybrid Recommendation ». In : *Proceedings of the 19th International Conference on Intelligent User Interfaces*. IUI '14. Haifa, Israel : ACM, p. 235-240. doi : 10.1145/2557500.2557542 (*citée page* 123).

Parsia, B., N. Matentzoglu, R. S. Gonçalves, B. Glimm et A. Steigmiller (2017). « The OWL Reasoner Evaluation (ORE) 2015 Competition Report ». In : *Journal of Automated Reasoning* 59.4, p. 455-482. doi : 10.1007/s10817-017-9406-8 (*citée page* 29).

Peng, W., M. O. Ward et E. A. Rundensteiner (2004). « Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering ». In : *INFOVIS '04 : Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'04)*. IEEE, p. 89-96. doi : 10.1109/infvis.2004.15 (*citée page* 91).

Perin, C., P. Dragicevic et J.-D. Fekete (2014). « Revisiting Bertin Matrices : New Interactions for Crafting Tabular Visualizations ». In : *IEEE Transactions on Visualization and Computer Graphics* 20.12, p. 2082-2091. doi : 10.1109/tvcg.2014.2346279 (*citée page* 87).

Picó, Y. (2012). « Chapter 2 - Data Analysis and Chemometrics ». In : *Chemical Analysis of Food : Techniques and Applications*. Boston : Academic Press, p. 25-57. doi : 10.1016/c2010-0-64808-5 (*citée page* 136).

Piepoli, M.-F., A.-W. Hoes, S. Agewall, C. Albus, C. Brotons, A.-L. Catapano et M.-T. Cooney (2016). « 2016 European Guidelines on cardiovascular disease prevention in clinical practice ». In : *European Heart Journal* 37.29, p. 2315-2381. doi : 10.1093/eurheartj/ehw106 (*citée page* 11, 14, 18).

Pinxteren, Y., G. Geleijnse et P. Kamsteeg (2011). « Deriving a recipe similarity measure for recommending healthful meals ». In : *Proceedings of the 15th international conference on Intelligent user interfaces - IUI '11*. ACM Press. doi : 10.1145/1943403.1943422 (*citée page* 112).

Pivk, A., P. Cimiano, Y. Sure, M. Gams, V. Rajkovič et R. Studer (2007). « Transforming arbitrary tables into logical form with TARTAR ». In : *Data & Knowledge Engineering* 60.3, p. 567-595. doi : 10.1016/j.datak.2006.04.002 (*citée page* 46).

Pletcher, M. J. et A. E. Moran (2017). « Cardiovascular Risk Assessment ». In : *Medical Clinics of North America* 101.4, p. 673-688. doi : 10.1016/j.mcna.2017.03.002 (*citée page* 17).

Powers, P. et L. Hoover (1989). « Calculating the nutrient composition of recipes with computers ». In : *Journal of the American Dietetic Association* 89.2, p. 224-232 (*citée page* 137).

Pravin, P. C., S. V. Patil et S. H. Dinde (2014). « Knowledge discovery from XML Database ». In : *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*. T. 1. IJIRAE, p. 172-178 (*citée page* 48).

Prud'hommeaux, E. et A. Seaborne (2008). *SPARQL Query Language for RDF*. W3C Working Draft. W3C (*citée page* 5, 29).

Pujahari, A. et V. Padmanabhan (2015). « Group Recommender Systems : Combining User-User and Item-Item Collaborative Filtering Techniques ». In : *2015 International Conference on Information Technology (ICIT)*. IEEE, p. 148-152. doi : 10.1109/icit.2015.36 (*citée page* 110).

Quinio, C., A. Biltoft-Jensen, S. D. Henauw, M. J. Gibney, I. Huybrechts, S. N. McCarthy, J. L. O'Neill, I. Tetens, A. Turrini et J.-L. Volatier (2007). « Comparison of different nutrient profiling schemes to a new reference method using dietary surveys ». In : *European Journal of Nutrition* 46.S2, p. 37-46. doi : 10.1007/s00394-007-2005-4 (*citée page* 128, 129).

Rafelsberger, W. (2013). « Interactive Visualization of Evolving Force-Directed Graphs ». In : *Design, User Experience, and Usability. Web, Mobile, and Product Design*. Springer Berlin Heidelberg, p. 553-559. doi : 10.1007/978-3-642-39253-5_61 (*citée page* 86).

Rajman, M. et R. Besançon (1998). « Text Mining - Knowledge extraction from unstructured textual data ». In : *Studies in Classification, Data Analysis, and Knowledge Organization*. Springer Berlin Heidelberg, p. 473-480. doi : 10.1007/978-3-642-72253-0_64 (*citée page* 44).

Ramadier, L. et M. Lafourcade (2016). « Patrons sémantiques pour l'extraction de relations entre termes - Application aux comptes rendus radiologiques ». In : *TALN 2016*. jep-taln2016. Paris, France (*citée page* 44).

Rashid, A.-M., I. Albert, D. Cosley, S. K. Lam, S. M. McNee, J. A. Konstan et J. Riedl (2002). « Getting to Know You : Learning New User Preferences in Recommender Systems ». In : *Proceedings of the 7th International Conference on Intelligent User Interfaces*. IUI '02. San Francisco, California, USA : ACM, p. 127-134. doi : 10.1145/502716.502737 (*citée page* 104).

Ravi, S. et M. Paşca (2008). « Using Structured Text for Large-scale Attribute Extraction ». In : *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. CIKM '08. Napa Valley, California, USA : ACM, p. 1183-1192. doi : 10.1145/1458082.1458238 (*citée page* 46).

Rayner, M., P. Scarborough et A. Kaur (2013). « Nutrient profiling and the regulation of marketing to children. Possibilities and pitfalls ». In : *Appetite* 62, p. 232-235. doi : 10.1016/j.appet.2012.06.021 (*citée page* 130).

Reinivuo, H., S. Bell et M.-L. Ovaskainen (2009). « Harmonisation of recipe calculation procedures in European food composition databases ». In : *Journal of Food Composition and Analysis* 22.5, p. 410-413. doi : 10.1016/j.jfca.2009.04.003 (*citée page* 138).

Resnick, P., N. Iacovou, M. Suchak, P. Bergstrom et J. Riedl (1994). « GroupLens : An Open Architecture for Collaborative Filtering of Netnews ». In : *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*. CSCW '94. Chapel Hill, North Carolina, USA : ACM, p. 175-186. doi : 10.1145/192844.192905 (*citée page* 110).

Ribeiro, R., F. Batista, J. P. Pardal, N. J. Mamede et H. S. Pinto (2006). « Cooking an Ontology ». In : *Artificial Intelligence : Methodology, Systems, and Applications*. Springer Berlin Heidelberg, p. 213-221. doi : 10.1007/11861461_23 (*citée page* 134).

Robertson, G. G. et J. D. Mackinlay (1993). « The Document Lens ». In : *Proceedings of the 6th Annual ACM Symposium on User Interface Software and Technology*. UIST '93. Atlanta, Georgia, USA : ACM, p. 101-108. doi : 10.1145/168642.168652 (*citée page* 83).

Roeser, S., R. Hillerbrand, P. Sandin et M. Peterson, éd. (2012). *Handbook of Risk Theory*. Springer Netherlands, p. 1187. doi : 10.1007/978-94-007-1433-5 (*citée page* 14, 15).

Roitman, H., Y. Messika, Y. Tsimerman et Y. Maman (2010). « Increasing Patient Safety Using Explanation-driven Personalized Content Recommendation ». In : *Proceedings of the 1st ACM International Health Informatics Symposium*. IHI '10. Arlington, Virginia, USA : ACM, p. 430-434. doi : 10.1145/1882992.1883057 (*citée page* 112).

Ronzano, F. et H. Saggion (2016). « Knowledge Extraction and Modeling from Scientific Publications ». In : *Semantics, Analytics, Visualization. Enhancing Scholarly Data*. Sous la dir. de G.-B. Alejandra, O. Francesco et P. Silvio. Cham : Springer International Publishing, p. 11-25 (*citée page* 49).

Roodenburg, A. J. C., B. M. Popkin et J. C. Seidell (2011). « Development of international criteria for a front of package food labelling system : the International Choices Programme ». In : *European Journal of Clinical Nutrition* 65.11, p. 1190-1200. doi : 10.1038/ejcn.2011.101 (*citée page* 129).

Rossi, F., Y. Lechevallier et A. E. Golli (2005). « Visualisation de la perception d'un site web par ses utilisateurs. » In : *Revue des Nouvelles Technologies de l'Information* Extraction et gestion des connaissances (EGC'2005), Actes des cinquièmes journées Extraction et Gestion des Connaissances, Paris, France, 18-21 janvier 2005, 2 Volumes, RNTI-E-3, p. 563-574 (*citée page* 97).

Rupp, M., J. Michaelis, D. McConnell et J. Smither (2016). « The Impact of Technological Trust and Self-Determined Motivation on Intentions to use Wearable Fitness Technology ». In : *Proceedings of the Human Factors and Ergonomics So-*

ciety Annual Meeting 60.1, p. 1434-1438. DOI : 10.1177/1541931213601329 (*citée page* 116).

SAINT-CYR, F. D. D. et D. PARADE (2018). « Représentation des Connaissances dans un Langage Visuel Typé : une Première Approche ». In : *29es Journées Francophones d'Ingénierie des Connaissances, IC 2018*. Sous la dir. de S. RANWEZ. AFIA. Nancy, France, p. 179-194 (*citée page* 78).

SALAS, P. E., E. MARX, A. MERA et J. VITERBO (2011). « RDB2RDF Plugin : Relational Databases to RDF Plugin for Eclipse ». In : *Proceedings of the 1st Workshop on Developing Tools As Plug-ins*. TOPI '11. Waikiki, Honolulu, HI, USA : ACM, p. 28-31. DOI : 10.1145/1984708.1984717 (*citée page* 48).

SAM, M., A. KRISNADHI, C. WANG, J. GALLAGHER et P. HITZLER (2014). « An Ontology Design Pattern for Cooking Recipes : Classroom Created ». In : *Proceedings of the 5th International Conference on Ontology and Semantic Web Patterns - Volume 1302*. WOP'14. Riva del Garda, Italy : CEUR-WS.org, p. 49-60 (*citée page* 134).

SAPORTA, G. (2011). *Probabilités, analyse des données et statistique*. Sous la dir. de T. (EDITIONS). 3ème édition révisée, p. 622 (*citée page* 40).

SARWAR, B., G. KARYPIS, J. KONSTAN et J. REIDL (2001). « Item-based Collaborative Filtering Recommendation Algorithms ». In : *Proceedings of the 10th International Conference on World Wide Web*. WWW '01. Hong Kong, Hong Kong : ACM, p. 285-295. DOI : 10.1145/371920.372071 (*citée page* 110).

SCHAFER, B., D. FRANKOWSKI, J. HERLOCKER et S. SEN (2007). « Collaborative Filtering Recommender Systems ». In : *The Adaptive Web*. Springer Berlin Heidelberg, p. 291-324. DOI : 10.1007/978-3-540-72079-9_9 (*citée page* 110, 111).

SCHAFER, J. B., J. KONSTAN et J. RIEDI (1999). « Recommender Systems in e-Commerce ». In : *Proceedings of the 1st ACM Conference on Electronic Commerce*. EC '99. Denver, Colorado, USA : ACM, p. 158-166. DOI : 10.1145/336992.337035 (*citée page* 104).

SCHEIN, A. I., A. POPESCUL, L. H. UNGAR et D. M. PENNOCK (2002). « Methods and Metrics for Cold-start Recommendations ». In : *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '02. Tampere, Finland : ACM, p. 253-260. DOI : 10.1145/564376.564421 (*citée page* 111).

SCHENK, S., C. SAATHOFF, S. STAAB et A. SCHERP (2009). « SemaPlorer—Interactive semantic exploration of data and media based on a federated cloud infrastructure ». In : *Web Semantics : Science, Services and Agents on the World Wide Web* 7.4, p. 298-304. DOI : 10.1016/j.websem.2009.09.006 (*citée page* 93).

SCHREIBER, G., H. AKKERMANS, A. ANJEWIERDEN, N. R. SHADBOLT, R. de HOOG, W. V. de VELDE et B. J. WIELINGA (2000). *Knowledge Engineering and Management - The CommonKADS Methodology*. MIT Press, p. 343 (*citée page* 68, 69).

SCHULZ, H. J. (2011). « Treevis.net : A Tree Visualization Reference ». In : *IEEE Computer Graphics and Applications* 31.6, p. 11-15. DOI : 10.1109/mcg.2011.103 (*citée page* 85).

SEDLMAIR, M., M. MEYER et T. MUNZNER (2012). « Design Study Methodology : Reflections from the Trenches and the Stacks ». In : *IEEE Transactions on Visualization and Computer Graphics* 18.12, p. 2431-2440. DOI : 10.1109/tvcg.2012.213 (*citée page* 123).

SEZGIN, E. et S. OZKAN (2013). « A systematic literature review on Health Recommender Systems ». In : *2013 E-Health and Bioengineering Conference (EHB)*. IEEE, p. 1-4. DOI : 10.1109/ehb.2013.6707249 (*citée page* 112).

Shannon, C. E. et W. Weaver (1949). *The mathematical theory of communication*. The Mathematical Theory of Communication. University of Illinois Press, p. 117 (*citée page* 66).

Sharko, J., G. Grinstein et K. Marx (2008). « Vectorized Radviz and Its Application to Multiple Cluster Datasets ». In : *IEEE Transactions on Visualization and Computer Graphics* 14.6, p. 1444-1427. doi : 10.1109/tvcg.2008.173 (*citée page* 91).

Shigarov, A. (2015). « Table understanding using a rule engine ». In : *Expert Systems with Applications* 42.2, p. 929-937. doi : 10.1016/j.eswa.2014.08.045 (*citée page* 52).

Shinzato, K. et K. Torisawa (2004a). « Acquiring Hyponymy Relations from Web Documents ». In : *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics : HLT-NAACL 2004*. HLT-NAACL (*citée page* 45).

— (2004b). « Extracting Hyponyms of Prespecified Hypernyms from Itemizations and Headings in Web Documents ». In : *Proceedings of the 20th International Conference on Computational Linguistics*. COLING '04. Geneva, Switzerland : Association for Computational Linguistics. doi : 10.3115/1220355.1220490 (*citée page* 45).

Shneiderman, B. (1992). « Tree visualization with tree-maps : 2-d space-filling approach ». In : *ACM Transactions on Graphics* 11.1, p. 92-99. doi : 10.1145/102377.115768 (*citée page* 85).

— (1996). « The eyes have it : A task by data type taxonomy for information visualizations ». In : *IN IEEE SYMPOSIUM ON VISUAL LANGUAGES*, p. 336-343 (*citée page* 90, 95, 97).

Silva, B. P. L. da, J.-F. Baget et M. Croitoru (2012). « A Generic Platform for Ontological Query Answering ». In : *Research and Development in Intelligent Systems XXIX*. Springer London, p. 151-164. doi : 10.1007/978-1-4471-4739-8_11 (*citée page* 92).

Snae, C. et M. Bruckner (2008). « FOODS : A Food-Oriented Ontology-Driven System ». In : *2008 2nd IEEE International Conference on Digital Ecosystems and Technologies*. IEEE, p. 168-176. doi : 10.1109/dest.2008.4635195 (*citée page* 33).

SNOMED-CT (2018). *SNOMED CT The Global Language of Healthcare. Dernière consultation : Septembre 2018*. url : https://www.snomed.org/snomed-ct (*citée page* 34).

Song, H., B. Kim, B. Lee et J. Seo (2010). « A Comparative Evaluation on Tree Visualization Methods for Hierarchical Structures with Large Fan-outs ». In : *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '10. Atlanta, Georgia, USA : ACM, p. 223-232. doi : 10.1145/1753326.1753359 (*citée page* 86).

Sporny, M., D. Longley, G. Kellogg, M. Lanthaler et N. Lindström (2018). *JSON-LD 1.1 A JSON-based Serialization for Linked Data*. W3C Working Draft. W3C (*citée page* 27).

Srinivasa, K., G. Siddesh et H. Srinidhi (2018). « Recommendation Systems ». In : *Computer Communications and Networks*. Springer International Publishing, p. 303-318. doi : 10.1007/978-3-319-77800-6_15 (*citée page* 104).

Stasko, J. et E. Zhang (2000). « Focus mathplus context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations ». In : *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*. IEEE Comput. Soc, p. 57-65. doi : 10.1109/infvis.2000.885091 (*citée page* 85).

Storey, M.-a., M. Musen, J. Silva, C. Best, N. Ernst, R. Fergerson et N. Noy (2001). « Jambalaya : Interactive visualization to enhance ontology authoring and knowledge acquisition in Protégé ». In : *Protégé. Workshop on Interactive Tools for Knowledge Capture (K-CAP-2001* (*citée page* 90).

Studer, R., R. V. Benjamins et D. Fensel (1998). « Knowledge Engineering : Principles and Methods ». In : *Data Knowl. Eng.* 25.1-2, p. 161-197. doi : 10.1016/S0169-023X(97)00056-6 (*citée page* 33).

Subramaniyaswamy, V., G. Manogaran, R. Logesh, V. Vijayakumar, N. Chilamkurti, D. Malathi et N. Senthilselvan (2018). « An ontology-driven personalized food recommendation in IoT-based healthcare system ». In : *The Journal of Supercomputing*, p. 1-33. doi : 10.1007/s11227-018-2331-8 (*citée page* 113, 118).

Sure, Y., M. Erdmann, J. Angele, S. Staab, R. Studer et D. Wenke (2002). « OntoEdit : Collaborative Ontology Development for the Semantic Web ». In : *The Semantic Web - ISWC 2002*. Springer Berlin Heidelberg, p. 221-235. doi : 10.1007/3-540-48005-6_18 (*citée page* 90).

Tarawaneh, R. M., P. Keller et A. Ebert (2012). « A General Introduction To Graph Visualization Techniques ». In : OpenAccess Series in Informatics (OASIcs) 27. Sous la dir. de C. Garth, A. Middel et H. Hagen, p. 151-164. doi : 10.4230/OASIcs.VLUDS.2011.151 (*citée page* 83, 85).

Teboul, J. et P. Damier (2017). *Le neuroleadership - Le cerveau face à la décision et au changement*. Editions Odile Jacob. Paris : Editions Odile Jacob, p. 340 (*citée page* 80).

Tekli, J., N. Charbel et R. Chbeir (2016). « Building semantic trees from XML documents ». In : *Web Semantics : Science, Services and Agents on the World Wide Web* 37-38, p. 1-24. doi : 10.1016/j.websem.2016.03.002 (*citée page* 48).

Teng, C.-Y., Y.-R. Lin et L. A. Adamic (2012). « Recipe Recommendation Using Ingredient Networks ». In : *Proceedings of the 4th Annual ACM Web Science Conference*. WebSci '12. Evanston, Illinois : ACM, p. 298-307. doi : 10.1145/2380718.2380757 (*citée page* 118).

Teufel, E. (2018). « Risk Factors for Cardiovascular Disease ». In : *Encyclopedia of Cardiovascular Research and Medicine*. Elsevier, p. 307-314. doi : 10.1016/b978-0-12-809657-4.99837-5 (*citée page* 10).

Teulier, R., J. Charlet et P. Tchounikine (2005). « L'ingénierie des connaissances : acquis et nouvelles perspectives ». In : *Ingénierie des connaissances*. L'Harmattan, p. 11-26 (*citée page* 5, 67).

Torre, G. L., V. Palmeri, L. Pagano, G. Nati, S. Sernia et A. Mannocci (2018). « Evaluation of cardiovascular risk profile : a comparative analysis between CUORE algorithm and the Framingham risk scores ». In : *European Review for Medical and Pharmacological Sciences* 22.8, p. 2398-2404. doi : 10.26355/eurrev_201804_14832 (*citée page* 18).

Tounsi, M., C. Lopez, C. Faron-Zucker, E. Cabrio, F. Gandon et F. Segond (2017). « Peuplement d'une base de connaissance par annotation automatique de textes relatifs à la cosmétique ». In : *28es Journées francophones d'Ingénierie des Connaissances IC 2017*. Sous la dir. de C. Roussey. Actes IC 2017 28es Journées francophones d'Ingénierie des Connaissances. Caen, France, p. 104-114 (*citée page* 44).

Truett, J., J. Cornfield et W. Kannel (1967). « A multivariate analysis of the risk of coronary heart disease in Framingham ». In : *Journal of Chronic Diseases* 20.7, p. 511-524. doi : 10.1016/0021-9681(67)90082-3 (*citée page* 18).

Tumnark, P., L. Oliveira et N. Santibutr (2013). « Ontology-Based Personalized Dietary Recommendation for Weightlifting ». In : *Proceedings of the 2013 International Workshop on Computer Science in Sports*. Atlantis Press. doi : 10.2991/iwcss-13.2013.13 (*citée page* 118).

Turo, D. (2003). « Hierarchical Visualization with Treemaps : Making Sense of Pro Basketball Data ». In : *The Craft of Information Visualization*. Elsevier, p. 237-238. doi : 10.1016/b978-155860915-0/50031-7 (*citée page* 90).

Ueda, M., S. Asanuma, Y. Miyawaki et S. Nakajima (2014). *Recipe Recommendation Method by Considering the User's Preference and Ingredient Quantity of Target Recipe*. S. I. Ao and Oscar Castillo and Craig Douglas and David Dagan Feng and Jeong-A Lee. Newswood Limited, p. 519-523 (*citée page* 112).

Ueda, M., Y. Morishita, T. Nakamura, N. Takata et S. Nakajima (2016). « A Recipe Recommendation System That Considers User's Mood ». In : *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services*. iiWAS '16. Singapore, Singapore : ACM, p. 472-476. doi : 10.1145/3011141.3011192 (*citée page* 113).

Unbehauen, J., S. Hellmann, S. Auer et C. Stadler (2012). « Knowledge Extraction from Structured Sources ». In : *Search Computing*. Springer Berlin Heidelberg, p. 34-52. doi : 10.1007/978-3-642-34213-4_3 (*citée page* 43, 48).

Unwin, I. D. (2000). « EUROFOODS Guidelines for Recipe Information Management ». In : *Journal of Food Composition and Analysis* 13.4, p. 745-754. doi : 10.1006/jfca.2000.0911 (*citée page* 137).

Ursyn, A. (2014). *Cognitive processes involved in visual thought*. IGI Global, p. 131-173. doi : 10.4018/978-1-4666-4703-9 (*citée page* 89).

— (2015). « Visualization as Communication with Graphic Representation ». In : *Encyclopedia of Information Science and Technology, Third Edition*. IGI Global, p. 2131-2139. doi : 10.4018/978-1-4666-5888-2.ch206 (*citée page* 89).

Vaillant, P. (1999). *Sémiotique des langages d'icônes*. Bibliothèque de grammaire et de linguistique. Honoré Champion, p. 304 (*citée page* 78).

Valdez, A. C., S. Bruns, C. Greven, U. Schroeder et M. Ziefle (2015). « What Do My Colleagues Know ? Dealing with Cognitive Complexity in Organizations Through Visualizations ». In : *Lecture Notes in Computer Science*. Springer International Publishing, p. 449-459. doi : 10.1007/978-3-319-20609-7_42 (*citée page* 123).

Vandenbussche, P.-Y. (2011). « Définition d'un cadre formel de représentation des Systèmes d'Organisation de la Connaissance ». Theses. Université Pierre et Marie Curie - Paris VI (*citée page* 32).

Vandervalk, B., L. Mccarthy et M. Wilkinson (2008). « CardioSHARE : Web Services for the Semantic Web ». In : *Semantic Web Challenge, (2008)* (*citée page* 36).

Vartiainen, E. (2018). « The North Karelia Project : Cardiovascular disease prevention in Finland ». In : *Global Cardiology Science and Practice* 2018.2. doi : 10.21542/gcsp.2018.13 (*citée page* 20).

Viot-Blanc, V. (2010). « Le manque de sommeil favorise-t-il l'obèsité, le diabète et les maladies cardiovasculaires ». In : *Mèdecine du Sommeil* 7.1, p. 15-22. doi : 10.1016/j.msom.2010.01.006 (*citée page* 11).

VOGEL, C. (1989). « Knowledge Oriented Design (KOD) : la mise en oeuvre ». In : sous la dir. d'E. MASSON (*citée page* 68).

VYTH, E. L., I. H. STEENHUIS, A. J. ROODENBURG, J. BRUG et J. C. SEIDELL (2010). « Front-of-pack nutrition label stimulates healthier product development : a quantitative analysis ». In : *International Journal of Behavioral Nutrition and Physical Activity* 7.1, p. 65. DOI : 10.1186/1479-5868-7-65 (*citée page* 130).

W., S. F., J. REPS, J. KAI, J. M. GARIBALDI et N. QURESHI (2017). « Can machine-learning improve cardiovascular risk prediction using routine clinical data ? » In : *PLOS ONE* 12.4, p. 14. DOI : 10.1371/journal.pone.0174944 (*citée page* 19).

WANG, T. D. et B. PARSIA (2006). « CropCircles : Topology Sensitive Visualization of OWL Class Hierarchies ». In : *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, p. 695-708. DOI : 10.1007/11926078_50 (*citée page* 91).

WARD, M. O., G. GRINSTEIN et D. KEIM (2015). *Interactive Data Visualization - Foundations, Techniques, and Applications, Second Edition*. 2 (29 mai 2015). Boca Raton London New York : A K Peters/CRC Press, p. 578 (*citée page* 82-86).

WEIMING, L., Z. ZHENYU, L. RENJIE, D. HAO, Y. SHANSONG et W. BAOGANG (2015). « Mining RDF from Tables in Chinese Encyclopedias ». In : *Natural Language Processing and Chinese Computing*. Springer International Publishing, p. 285-298. DOI : 10.1007/978-3-319-25207-0_24 (*citée page* 46).

WESTENBRINK, S., M. ROE, M. OSEREDCZUK, I. CASTANHEIRA et P. FINGLAS (2016). « EuroFIR quality approach for managing food composition data where are we in 2014 ? » In : *Food Chemistry* 193, p. 69-74. DOI : 10.1016/j.foodchem.2015.02.110 (*citée page* 137).

WHO (2007). *Prevention of cardiovascular disease : pocket guidelines for assessment and management of cardiovascular risk : (WHO/ISH cardiovascular risk prediction charts for the European Region)*. technical meeting. London, United Kingdom (*citée page* 10, 15, 18).

— (2011). *Nutrient profiling : report of a technical meeting*. technical meeting. London, United Kingdom (*citée page* 129, 130).

— (2017). *World Health Statistics 2017 :Monitoring Health for the SDGs Sustainable Development Goals*. World Health Statistics Annual. World Health Organization (*citée page* 5, 10).

— (2018). *71e Assemblée mondiale de la Santé – principaux débats du 24 mai*. World Health Statistics Annual. World Health Organization (*citée page* 4).

WIESNER, M. et D. PFEIFER (2014). « Health Recommender Systems : Concepts, Requirements, Technical Basics and Challenges ». In : *International Journal of Environmental Research and Public Health* 11.3, p. 2580-2607. DOI : 10.3390/ijerph110302580 (*citée page* 112).

« High Risk (Prevention) Strategy » (2008). In : sous la dir. de K. WILHELM, p. 675-675. DOI : https://doi.org/10.1007/978-1-4020-5614-7 (*citée page* 16, 20).

WILHELMSEN, L. (2017). « Cardiovascular Disease Prevention ». In : *International Encyclopedia of Public Health*. Elsevier, p. 438-447. DOI : 10.1016/b978-0-12-803678-5.00055-2 (*citée page* 19).

WILLIAMSON, D. A. (2009). « Impact of a Weight Management Program on Health-Related Quality of Life in Overweight Adults With Type 2 Diabetes ». In : *Archives of Internal Medicine* 169.2, p. 163. DOI : 10.1001/archinternmed.2008.544 (*citée page* 120).

Winstead, J. et N. Gershon (1996). « The World-Wide Web : A Global Source of Data and Information ». In : *Data and Knowledge in a Changing World*. Springer Berlin Heidelberg, p. 21-28. doi : 10.1007/978-3-642-85248-0_3 (*citée page* 22).

Wolfgang, K.-H., K. Sigbert et R. Bernd (2015). *Introduction to Statistics : Using Interactive MM Stat Elements*. Springer International Publishing, p. 516. doi : 10.1007/978-3-319-17704-5 (*citée page* 40).

Woodward, M., P. Brindle et H. T.-P. and (2005). « Adding social deprivation and family history to cardiovascular risk assessment : the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC) ». In : *Heart* 93.2, p. 172-176. doi : 10.1136/hrt.2006.108167 (*citée page* 18).

Wu, W., L. He et J. Yang (2012). « Evaluating recommender systems ». In : *Seventh International Conference on Digital Information Management (ICDIM 2012)*. IEEE, p. 56-61. doi : 10.1109/icdim.2012.6360092 (*citée page* 116).

Yang, L., C. K. Hsieh, H. Yang, N. Dell, S. J. Belongie et D. Estrin (2017). « Yum-Me : A Personalized Nutrient-Based Meal Recommender System ». In : *ACM Trans. Inf. Syst.* 36.1, p. 1-31. doi : 10.1145/3072614 (*citée page* 113).

Yang, L., C.-K. Hsieh, H. Yang, J. P. Pollak, N. Dell, S. Belongie, C. Cole et D. Estrin (2017). « Yum-Me ». In : *ACM Transactions on Information Systems* 36.1, p. 1-31. doi : 10.1145/3072614 (*citée page* 118).

Yeon-Seok, K. et L. Kyong-Ho (2008). « Extracting logical structures from HTML tables ». In : *Computer Standards & Interfaces* 30.5, p. 296-308. doi : 10.1016/j.csi.2007.08.006 (*citée page* 42).

Yoshinaga, N. et K. Torisawa (2007). « Open-Domain Attribute-Value Acquisition from Semi-Structured Texts ». In : *Proceedings of the 6th International Semantic Web Conference (ISWC-07), Workshop on Text to Knowledge : The Lexicon/Ontology Interface (OntoLex-2007)*, p. 55-66 (*citée page* 45).

Zaveri, A., J. Lehmann, S. Auer, M. Hassan, M. Sherif et M. Martin (2013). « Publishing and interlinking the Global Health Observatory dataset ». In : *Semantic Web* 4.3, p. 315-322. doi : 10.3233/SW-130102 (*citée page* 37).

Zazkis, R., E. Dubinsky et J. Dautermann (1996). « Coordinating Visual and Analytic Strategies : A Study of Students' Understanding of the Group D 4 ». In : *Journal for Research in Mathematics Education* 27.4, p. 435. doi : 10.2307/749876 (*citée page* 81).

Ziefle, M., C. Rocker et A. Holzinger (2011). « Medical Technology in Smart Homes : Exploring the User's Perspective on Privacy, Intimacy and Trust ». In : *2011 IEEE 35th Annual Computer Software and Applications Conference Workshops*. IEEE, p. 410-415. doi : 10.1109/compsacw.2011.75 (*citée page* 116).

Zweigenbaum, P., F. Hadouche et N. Grabar (2003). « Apprentissage de relations morphologiques en corpus ». In : *Actes de la 10ème conférence sur le Traitement Automatique des Langues Naturelles*. Batz-sur-Mer, France : Association pour le Traitement Automatique des Langues, p. 285-294 (*citée page* 44).

Annexe A

Evolution de l'interface de l'outil MCVGraphViz

La figure A.1 présente la première version de l'outil MCVGraphViZ : (1) aperçu de l'interface de visualisation des interactions entre les facteurs de risque cardiovasculaire affiché sous forme de nœuds et d'arcs dans la vue initiale ; (2) vue focalisée obtenue *via* la sélection d'un nœud ; (3) zone de description du nœud sélectionné ; (4) historique des graphes visité dans le parcours actuel.

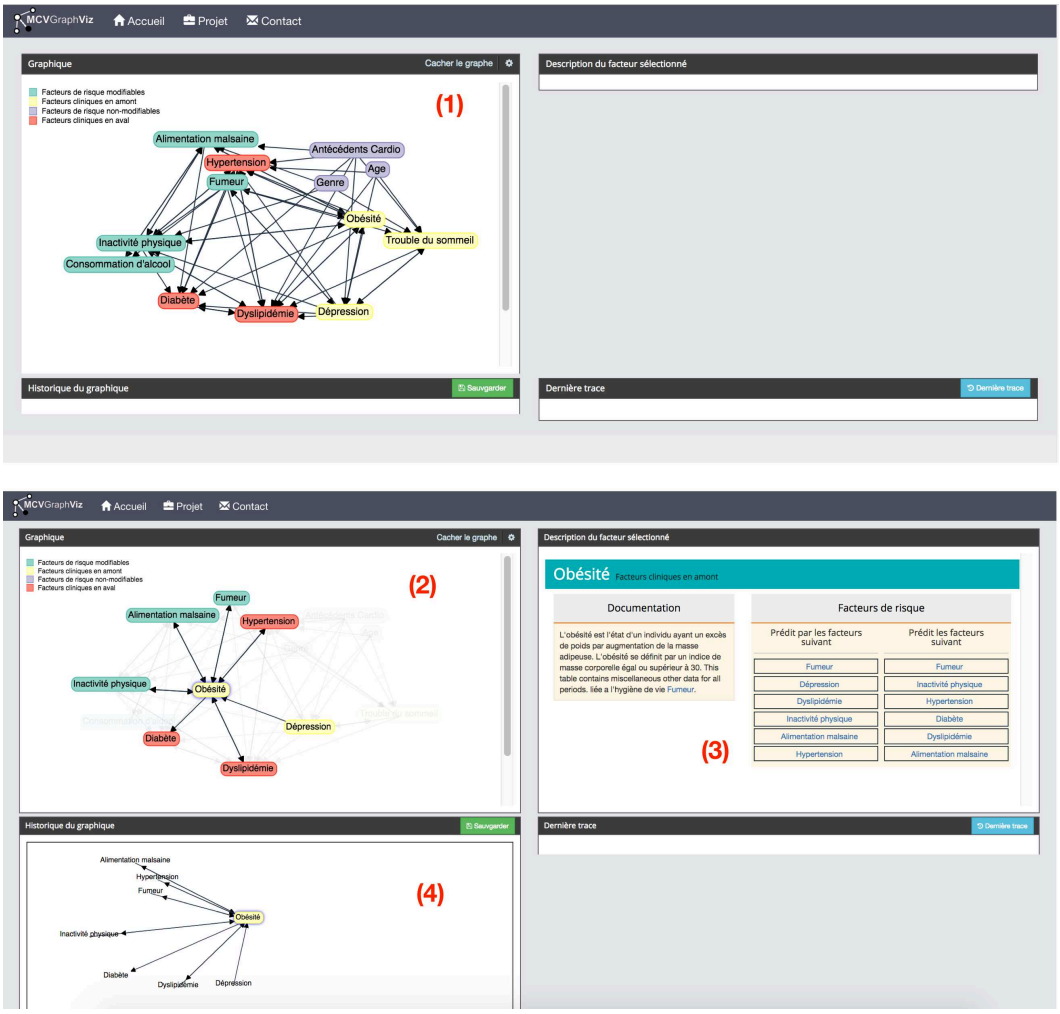


FIGURE A.1 – Première version de l'outil MCVGraphViz

La figure A.2 présente la seconde version de l'outil MCVGraphViZ : (1) aperçu de l'interface de visualisation des interactions entre les facteurs de risque cardiovasculaire, affiché sous forme de nœuds et d'arcs dans la vue initiale. Les liens entre les facteurs de risque sont étiquetés ; (2) zone de description des noeuds avec un accès à de la documentation issue de la HAS ; (3) vue focalisée obtenue *via* la sélection d'un nœud. (4) description du nœud sélectionné.

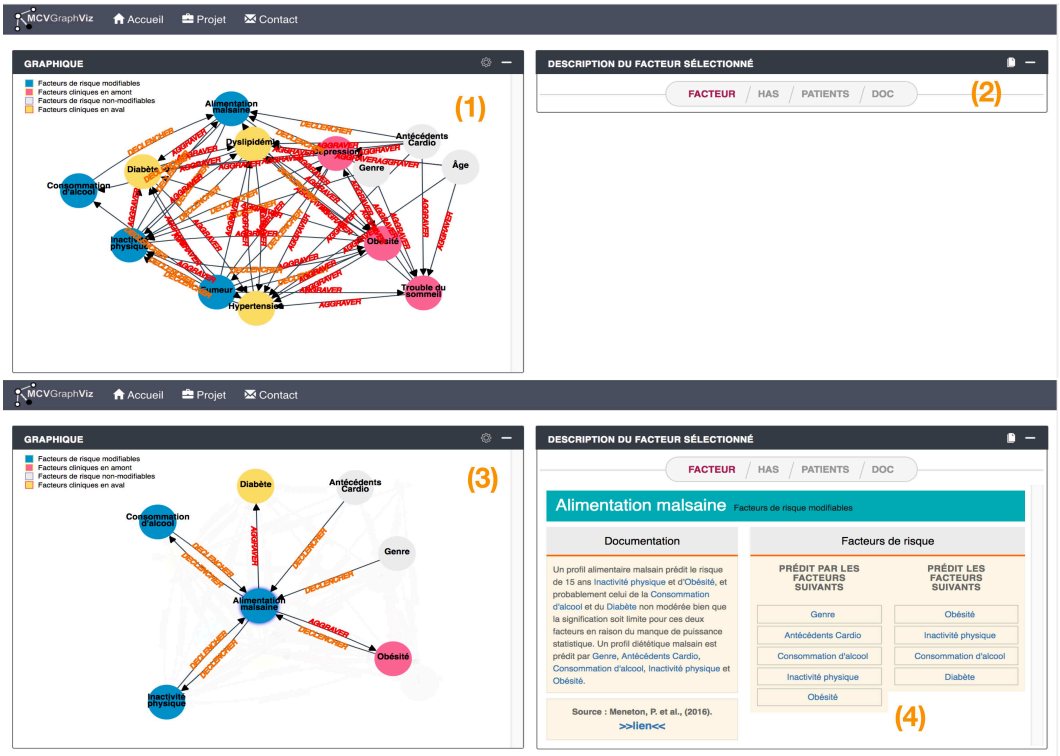


FIGURE A.2 – Seconde version de l'outil MCVGraphViz

La figure A.3 présente la troisième version de l'outil MCVGraphViZ : (1) aperçu de l'interface de visualisation des interactions entre les facteurs de risque cardiovasculaire, affiché sous forme de nœuds avec des diamètres variés selon le nombre de nœuds connectés avec celui-ci. Les liens entre les facteurs de risque sont étiquetés et l'épaisseur des relations est proportionnelle à la probabilité de la relation. Un moteur de recherche dans le graphe est intégré ; (2) vue focalisée obtenue *via* la sélection d'un nœud ; (3) historique des graphes visité dans le parcours actuel (4) zone de description des noeuds avec un accès à de la documentation issue de la HAS et un système d'accès a des profils de patients évalués selon le système SCORE de la HAS avec un ensemble de recommandations associées selon ce SCORE.

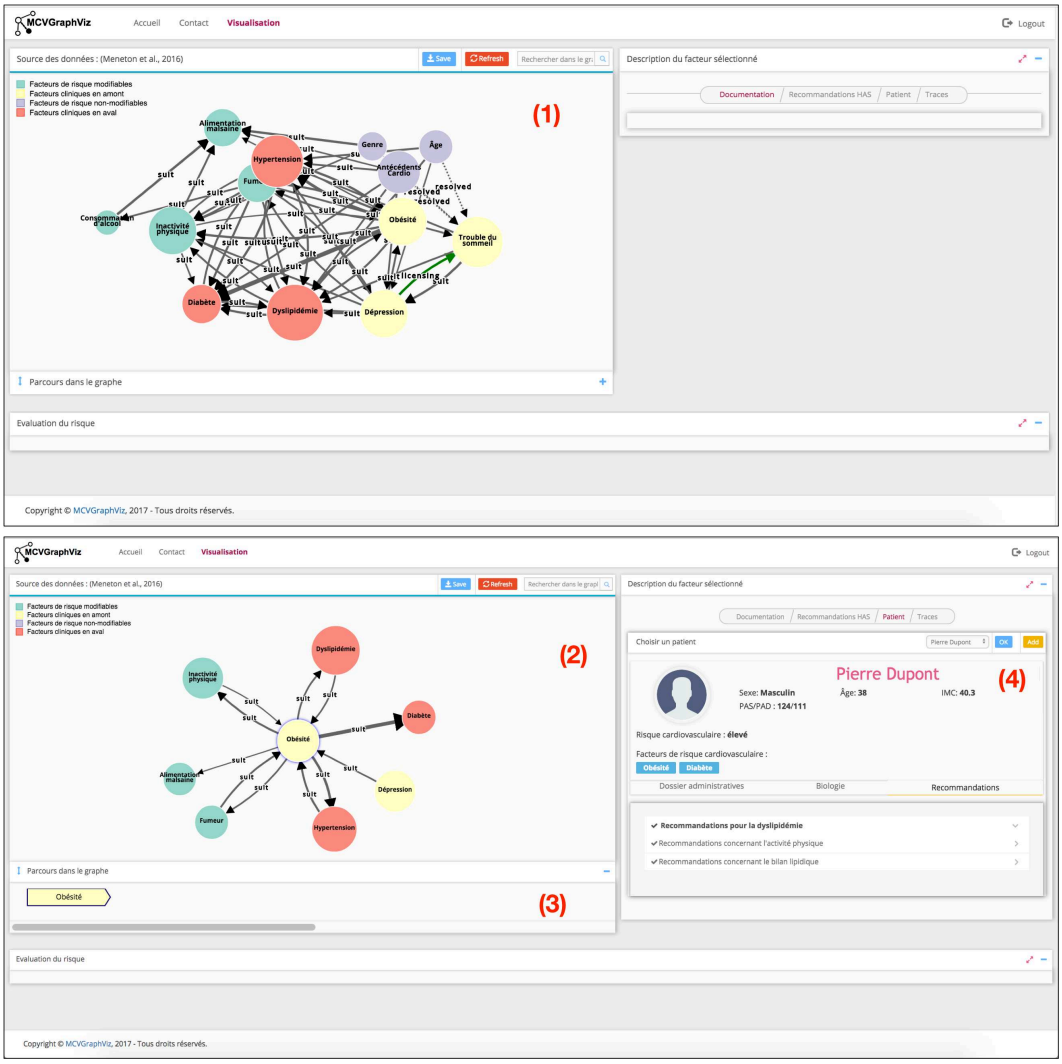


FIGURE A.3 – Troisième version de l'outil MCVGraphViz

Annexe B

Qualification nutritionnelle des recettes

La figure B.1 présente le scénario dans lequel la qualification de la recette doit être réalisée. Ce scénario nous a permis de mieux identifier les informations à fournir à l'utilisateur et pour quel objectif.

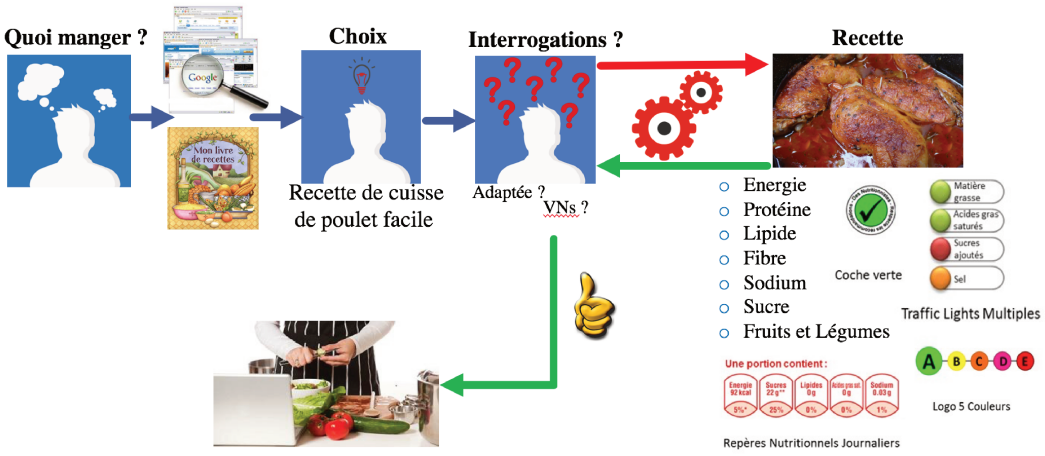


FIGURE B.1 – Scénario dans lequel la qualification de la recette doit être réalisée

La figure B.2 présente l'approche adoptée pour annoter à l'aide de la RTO chaque recette (ingrédient et préparation). Elle permet ainsi d'identifier le produit de base de chaque ingrédient dans la liste d'ingrédients et dans chaque instruction de préparation.

La figure B.3 présente l'approche d'appariement de l'ingrédient de la recette avec celui de la base de données Nutrinet. Elle consiste à récupérer les valeurs nutritionnelles pour chaque ingrédient en identifiant sa correspondance dans la base de données.

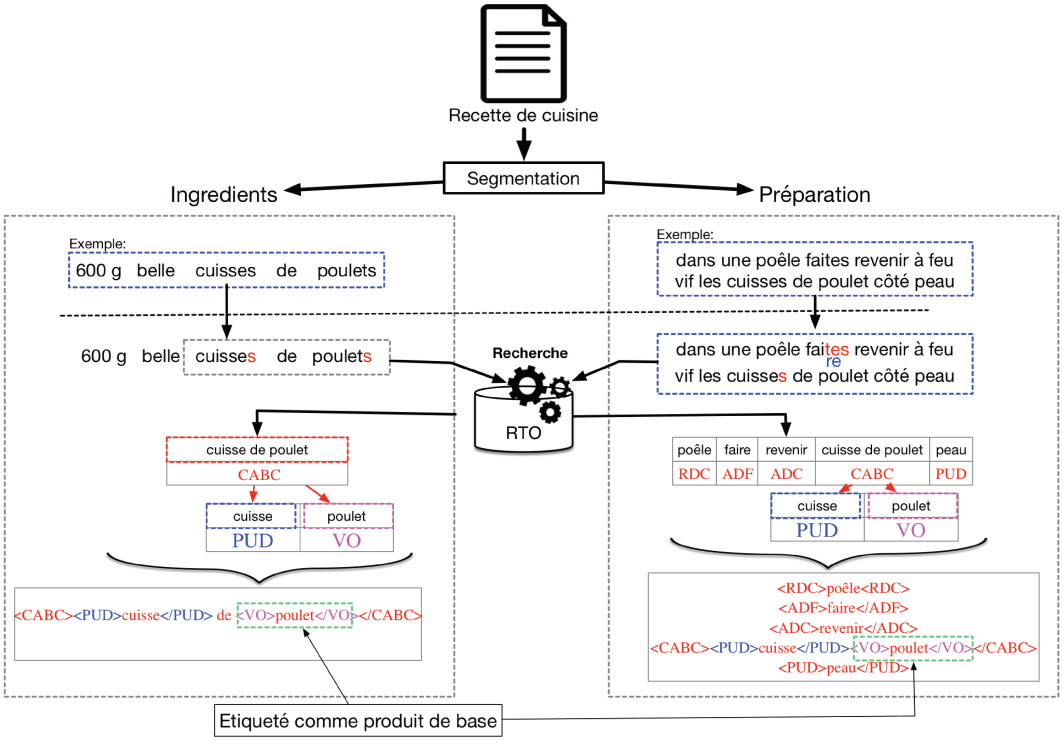


FIGURE B.2 – Démarche d'annotation des recettes de cuisine à l'aide de la RTO

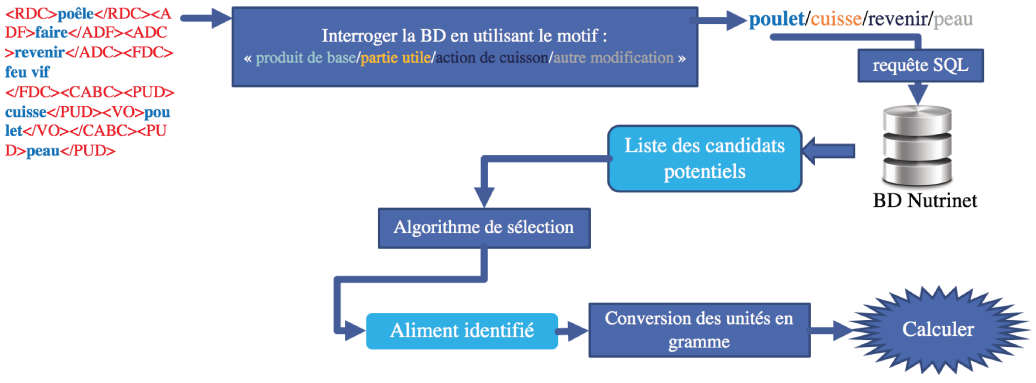


FIGURE B.3 – Appariement de l'ingrédient de la recette avec celui de la base de données Nutrinet

Table des matières

Sommaire ..... **vii**

Préambule **1**

Introduction **3**

Chapitre 1 Contexte applicatif et scientifique **9**
1.1 Projet d'amélioration de la prévention du risque cardiovasculaire 10
1.2 Epidémiologie des maladies cardiovasculaires 10
1.2.1 Généralités sur les maladies cardiovasculaires 10
1.2.2 Statistiques des maladies cardiovasculaires 10
1.2.3 Les facteurs associés aux maladies cardiovasculaires 11
1.3 Prévention du risque 14
1.3.1 Définition du risque 14
1.3.2 Définition de la prévention 15
1.3.3 Systèmes d'évaluation du risque 16
1.3.4 Approches de prévention sensibles au domaine 20
1.3.5 Vers un système de prévention du risque cardiovasculaire informatisé 21
1.4 Le web sémantique 21
1.4.1 Présentation générale du web sémantique 21
1.4.2 Les standards du W3C 24
1.5 Outils du web sémantique pour l'organisation de la connaissance en e-santé 30
1.5.1 Thésaurus et ontologies 31
1.5.2 Référentiels de données sémantiques 36
1.6 Conclusion partielle 38

Chapitre 2 Exploitation des présentations de résultats statistiques **39**
2.1 Définition et terminologie 40
2.1.1 Statistiques 40
2.1.2 Démarche statistique 40
2.1.3 Présentation des résultats statistiques 41
2.1.4 Discussion 42
2.2 Approches d'extraction de connaissances à partir de textes 43
2.2.1 Approche sur textes non structurés 43
2.2.2 Approche sur textes structurés 45
2.2.3 Discussion et positionnement 50

2.3 Approche proposée 50
2.3.1 Conversion des documents de présentation de résultats statistique 50
2.3.2 Localisation et extraction de l'information 52
2.3.3 Construction des triplets 54
2.4 Expérimentation 56
2.4.1 Domaine des maladies cardiovasculaires 56
2.4.2 Domaine financier 59
2.5 Evaluation de l'approche 60
2.6 Conclusion 64

Chapitre 3 Modélisation des connaissances 65
3.1 Définition et objectifs 66
3.1.1 La notion de connaissance 66
3.1.2 Ingénierie des connaissances 66
3.1.3 La modèlisation des connaissances 67
3.1.4 Objectif 68
3.2 Approche proposée 69
3.2.1 E1 : construction du Premier modèle conceptuel 70
3.2.2 E2 : définition du modèle conceptuel complet 71
3.2.3 Discussion 74
3.2.4 E3 : définition et validation du modèle conceptuel complet 74
3.3 Conclusion 76

Chapitre 4 MCVGraphViz, un outil de visualisation 77
4.1 Définition et objectifs 78
4.1.1 Cognition et perception visuelle 78
4.2 Visualisation d'information *vs* connaissances 81
4.2.1 Définition de la visualisation 81
4.2.2 Visualisation d'information 82
4.2.3 Visualisation de connaissances 88
4.2.4 Quelle approche choisir pour le graphe de connaissances ? 94
4.3 Conception de MCVGraphViz 95
4.3.1 Scénario d'usage 95
4.3.2 Encodage graphique 95
4.3.3 Aspect dynamique 96
4.3.4 Implémentation 97
4.4 Conclusion 101

Chapitre 5 Systèmes de recommandation 103
5.1 Introduction 104
5.2 Les systèmes de recommandation 104
5.2.1 Définitions 104
5.2.2 Classification des systèmes de recommandation 105
5.3 Systèmes de recommandation en santé 112
5.3.1 Composantes des systèmes de recommandation en santé 113
5.3.2 Comparaison des systèmes de recommandation en santé 117
5.4 Approche proposée 119
5.4.1 Utilisateurs du système 119
5.4.2 Méthode de recommandation 120
5.5 Conclusion 126

Chapitre 6 Qualification nutritionnelle des recettes de cuisine 127
6.1 Introduction 128
6.2 Systèmes de profilage nutritionnel 129
6.2.1 Définition 129
6.2.2 Eléments de base des systèmes de profilage nutritionnel 130
6.3 Qualification nutritionnelle des recettes de cuisine 131
6.3.1 Définition 131
6.3.2 Connaissances du domaine 133
6.3.3 Tables et base de données sur la composition des aliments 134
6.3.4 Calcul des valeurs nutritionnelles des recettes 136
6.3.5 Traduction visuelle des valeurs nutritionnelles 137
6.4 Approche proposée 138
6.4.1 Description des ressources utilisées 138
6.4.2 Description des étapes de la méthodologie 143
6.4.3 Enrichissement lexical des produits composant les ingrédients 144
6.4.4 Génération du fichier de calcul 147
6.4.5 Calcul et attribution du score 148
6.4.6 Traduction du score sur l'échelle graphique 150
6.5 Evaluation des résultats 150
6.6 Conclusion 153

Conclusion et perspectives 155

Bibliographie 186

Chapitre A Evolution de l'interface de l'outil MCVGraphViz 187

Chapitre B Qualification nutritionnelle des recettes 191

Table des matières 195

Index des tableaux 197

Index des figures 201

Liste des tableaux

1 Tableau synoptique du document 1

1.1 Liste des facteurs de risque cardiovasculaire 12
1.2 Suite de la liste des facteurs de risque cardiovasculaire 13

2.1 Tableau comparatif des outils de conversion de documents PDF 52
2.2 Résultats expérimentaux 62

4.1 Propriétés des différentes variables selon (Bertin, 1999). Pour les propriétés, « + » indique que la propriété est présente pour une variable, « ++ » qu'elle est très prononcée, « (+) » qu'elle peut être présente mais à certaines conditions. 80

5.1 Exemple d'une matrice de notes 105
5.2 Comparaison de quelques systèmes de recommandation 118

6.1 Comparaisons de quatre valeurs nutritionnelles de deux aliments de Nutrinet et USDA 140
6.2 Valeurs nutritionnelles de la Tomate fraîche et cuite 147
6.3 Résultats pour les six premières recettes 151

Table des figures

1 Caractéristiques générales des travaux de la thèse 1
2 Guide de lecture du manuscrit 2

1.1 Transition démographique des maladies chroniques 10
1.2 Classification de stratégies préventives 15
1.3 Evolution du web (source : `https://c-marketing.eu/du-web-1-0-au-web-4-0/`) 22
1.4 Semantic Web Layer Cake ou la pile des standards du web sémantique (Gandon *et al.*, 2012) 25
1.5 Exemple de triplet RDF 26
1.6 Exemples de syntaxes RDF : la syntaxe RDF/XML, la syntaxe N3 et la syntaxe JSON-LD 27
1.7 Exemples de requête SPARQL 29
1.8 Nuage du web de données (source : `https://lod-cloud.net/`) 31
1.9 Interface de CardioShare 37

2.1 Démarche d'analyse d'un tableau statistique 41
2.2 Exemple de tableau simple et complexe 42
2.3 Différentes approches d'extraction des connaissances à partir de texte ... 44
2.4 Exemple de document XML décrivant une recette de cuisine 47
2.5 Etapes de la méthodologie proposée 51
2.6 Sortie obtenue avec PdftohtmlEX et PDFX 53
2.7 Stratégie d'extraction des triplets à partir des tableaux 55
2.8 Stratégie d'extraction de l'information appliquées sur le tableau 56
2.9 Modèle de triplets pour l'extraction d'information 57
2.10 Vocabulaire contrôlé construit 58
2.11 Exemple de tags réalisé sur un tableau à l'aide du vocabulaire contrôlé.. 58
2.12 Exemple de tableau extrait à partir du document HTML 59
2.13 Tableau (A) de la figure 2.12 après nettoyage 59
2.14 Triplets obtenus à partir du tableau (A) de la figure 2.12 60
2.15 Extraction d'un tableau figurant dans l'enquête 60
2.16 Liste des triplets obtenus à partir du tableau de la figure 2.15(B) 61
2.17 Résultats obtenus pour le facteur de risque « Fumeur » pour les jeux de données (D1) et (D3) 63
2.18 Résultats obtenus pour le facteur de risque « Obesity » pour les jeux de données (D1) et (D3) 63

3.1 Les trois axes du triangle sémiotique, la conjonction S3 (Le Moigne, 1990) 66
3.2 Les trois axes du triangle sémiotique selon (Aries *et al.*, 2008) 67

3.3 Modèle conceptuel de connaissance selon la notion de point de vue 70
3.4 Approche de construction du modèle conceptuel 70
3.5 Modèle conceptuel 71
3.6 Schéma du modèle conceptuel des interactions entre les facteurs de risque cardiovasculaire 72
3.7 Interface de l'outil MCVGraphViz avec le graphe de l'étude de (Meneton *et al.*, 2016) 73
3.8 Vue du facteur de risque « Trouble du sommeil » dans MCVGraphViz lors de la validation par l'expert 74
3.9 Vue du facteur « Fumeur » dans MCVGraphViz à partir de (D1) et (D2) 74
3.10 Nature des liens entre « Fumeur » et « diabète » suivant la présence ou non chez un individu 75

4.1 Exemples d'application de la loi de *Gestalt*. 79
4.2 Les variables retiniennes de (Bertin, 1999). 80
4.3 Modèle de (Card *et al.*, 1999) pour la visualisation d'information. 82
4.4 Structure d'arbre 84
4.5 Capture d'écran représentant nombre des techniques de visualisation d'arbres à cette date (source : `https://keshif.me/demo/treevis.html)` 84
4.6 Exemple de remplissage d'espace disponible (Ward *et al.*, 2015)..... 85
4.7 Exemple de représentation exploitant la structure arborescente des donneés 86
4.8 Exemple de graphe dirigé par la force : (a) issue de (Ward *et al.*, 2015) ; (b) issue de (Rafelsberger, 2013) ; (c) issue de (Paananen et Wong, 2009). 86
4.9 Exemple de graphe représenté par une matrice : (a) d'après (H. Li *et al.*, 2009) et (b) d'après (Kerzner *et al.*, 2017). 87
4.10 Exemple de carte de connaissances (source : `https://imindmap.com/gallery/`). 90
4.11 Techniques de visualisation exploitant la sémantique hiérarchique..... 90
4.12 Exemple de visualisation exploitant la sémantique hiérarchique 91
4.13 Exemple de plugin de visualisation pour Protégé..... 92
4.14 Exemple d'outil de visualisation exploitant la sémantique relationnelle .. 92
4.15 Exemple d'outil de visualisation exploitant la sémantique fondée sur les entités 93
4.16 Approche de visualisation dans MCVGraphViz 94
4.17 Liste des variables graphiques employées dans MCVGraphViz 96
4.18 Processus d'interaction entre : D3.js, DOM et l'interface de visualisation. 98
4.19 Interface de l'outil MCVGraphViz..... 99
4.20 Exemple d'action intégrée dans MCVGraphViz..... 99
4.21 Exemple d'informations contenu dans la zone (b). 100
4.22 Exemple de parcours dans le graphe 100
4.23 Interface de MCVGraphViz actualisée après la sélection d'un facteur de risque 101

5.1 Une architecture haut niveau d'un système de recommandation fondé sur le contenu (Lops *et al.*, 2010) 106
5.2 Conseils essentiels du Programme national nutrition-santé sur l'alimentation et l'activité physique (source : `http://eduscol.education.fr/cid47662/decouvrir-alimentation.html`) 121
5.3 Intégration des recommandations dans l'outil MCVGraphViz : accès utilisateur simple ou patient 124

5.4 Outil de qualification nutritionnelle des recettes de cuisine 125
5.5 Intégration des recommandations dans l'outil MCVGraphViz : accès médecin 125

6.1 Exemple de système de représentation graphique : (A) le système Feux tricolores, créé en 2013 au Royaume-Uni ; (B) le système GDA (Guideline Daily Amounts), créé en 2006 et reconnu en 2011 en Europe ; (C) le système SENS, créé en 2015 en France ; (D) le Nutri-Score créé en 2013 en France. 131
6.2 Structure d'une recette de cuisine 132
6.3 La recette de cuisine ratatouille décrite à l'aide du microformat hRecipe 133
6.4 Structure d'une recette de cuisine 139
6.5 Calcul du pourcentage d'apparition des ingrédients dans la préparation de la recette 140
6.6 Table de composition des aliments nutrinet 142
6.7 Modèle des aliments issu de (Despres, 2014) 142
6.8 Etapes de la méthodologie de calcul des valeurs nutitionnelles des recettes 143
6.9 Extrait de la structure du lexique construit à partir de la RTO 144
6.10 Axes de recherche pour l'enrichissement sémantique des tokens 145
6.11 Recherche de cuisse de poulet dans la RTO 146
6.12 Echelle de qualité nutritionnelle du PNNS 150
6.13 Comparaison sur la recette ratatouille 152

A.1 Première version de l'outil MCVGraphViz 187
A.2 Seconde version de l'outil MCVGraphViz 188
A.3 Troisième version de l'outil MCVGraphViz 189

B.1 Scénario dans lequel la qualification de la recette doit être réalisée 191
B.2 Démarche d'annotation des recettes de cuisine à l'aide de la RTO 192
B.3 Appariement de l'ingrédient de la recette avec celui de la base de données Nutrinet 192

Apport du web sémantique pour améliorer la prévention des maladies cardiovasculaires

Résumé : Les maladies cardiovasculaires (MCVs) et les cancers constituent la première et la seconde cause de mortalité et de morbidité en France chez les hommes et les femmes, et leur coût annuel est important. La prévention de ces maladies chroniques constitue, avec leur dépistage précoce et leur prise en charge rapide et efficace, un moyen possible de réduire ce coût. Cette prévention passe par l'identification des facteurs de risque modifiables associés à ces maladies chroniques, au rang desquels on retrouve l'alimentation et l'inactivité physique. Ainsi, le Programme National Nutrition Santé (PNNS) a été mis en place en France pour aider les Français à avoir une meilleure alimentation, afin de contribuer à réduire l'incidence de ces maladies. L'objectif de cette thèse est la construction d'un système de suggestions personnalisées fondé sur le profil de l'individu et son risque cardiovasculaire. Cette approche nécessite la mise en place d'une démarche interdisciplinaire faisant collaborer des chercheurs dans les domaines de l'informatique, de l'épidémiologie et de la nutrition. L'importance de cette collaboration se justifie par le besoin de produire des suggestions étayées par des recherches attestées dans ces domaines. Le premier apport de cette thèse est l'intégration des technologies du web sémantique dans une nouvelle approche d'évaluation du risque cardiovasculaire qui prend en compte les interactions entre ces facteurs. L'approche proposée consiste à : (i) extraire et exploiter les connaissances à partir des présentations statistiques des résultats ; (ii) construire un modèle conceptuel à partir de ces connaissances ; (iii) visualiser dynamiquement le modèle obtenu pour mieux comprendre les effets en cascade. La création de l'outil de visualisation MCVGraphViz a permis de mettre en œuvre cette stratégie. Le deuxième apport consiste à proposer une solution pour exploiter les connaissances présentes dans les plans de santé et les recommandations concernant la prévention des maladies cardiovasculaires en France. Ainsi, nous avons opté pour une approche modulaire intégrée dans l'outil MCVGraphViz qui permet de produire des recommandations (alimentation, activité physique, etc.) fondées sur le risque cardiovasculaire évalué et le profil de l'individu (préférences sensorielles, contraintes allergiques, capacité physique, etc.). Le troisième apport concerne la qualification nutritionnelle des recettes de cuisine pour un meilleur suivi : l'approche s'appuie sur des techniques de traitement automatique de la langue et des raisonnements ontologiques pour qualifier d'un point de vue nutritionnel des recettes de cuisine. De nombreuses perspectives sont exposées. La plupart d'entre elles visent à améliorer les systèmes de recommandation et l'expressivité de la base de connaissances sur les maladies cardiovasculaires.

Mots-clefs : Ingénierie des connaissances, Web sémantique, Services web, Visualisation de connaissances, Facteurs de risque et interactions, Maladies cardiovasculaires.

Abstract : Cardiovascular disease (CVD) and cancer are the leading cause of death and morbidity in men and women in France, and their annual cost is high. The prevention of these chronic diseases is, with their early detection and their rapid and effective management, a possible way to reduce this cost. This prevention involves identifying the modifiable risk factors associated with these chronic diseases, among which are diet and physical inactivity. Thus, the National Health Nutrition Program was set up in France, helping the French to have a better diet, to help reduce the incidence of these diseases. The objective of this thesis is the construction of a system of personalized suggestions based on the profile of the individual and his cardiovascular risk. This approach requires the establishment of an interdisciplinary approach involving researchers in the fields of computer science, epidemiology and nutrition. The importance of this collaboration is justified by the need to produce suggestions supported by proven research in these areas. The first contribution of this thesis is the integration of semantic web technologies into a new approach to cardiovascular risk assessment that takes into account the interactions between these factors. The proposed approach consists of: (i) extracting and exploiting knowledge from statistical presentations of results; (ii) construct a conceptual model based on this knowledge; (iii) dynamically visualize the resulting model to better understand the cascading effects. The creation of the visualization tool MCVGraphViz allowed to implement this strategy. The second contribution consists in proposing a solution to exploit the knowledge present in the health plans and the recommendations concerning the prevention of cardiovascular diseases in France. Thus, we opted for a modular approach integrated into the tool MCVGraphViz that allows to produce recommendations (diet, physical activity, etc.) based on the assessed cardiovascular risk and the profile of the individual (sensory preferences, allergic constraints, physical capacity, etc.). The third contribution concerns the nutritional qualification of cooking recipes for a better follow-up: the approach is based on techniques of automatic language processing and ontological reasoning to qualify a nutritional point of view of cooking recipes. Many perspectives are exposed. Most of them aim to improve referral systems and the expressivity of the cardiovascular disease knowledge base.

Keywords : Knowledge Engineering, Semantic Web, Web Services, Visualization of Knowledge, Risk Factors and Interactions, Cardiovascular Diseases.

DISCIPLINE : INFORMATIQUE

Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé (LIMICS) UMR_S 1142, 74 Rue de Marcel Cachin, 93017 Bobigny, France