NutriKG – Un Graphe de Connaissances pour Modéliser les Préférences et les Besoins Nutritionnels

Alexandre Combeau[1,2], Fatiha Saïs [2], Nageeta Kumari[3], Stéphane Dervaux [1], Cristina Manfredotti [1], Vincent Guigue [1], Paolo Viappiani [4]

[1] Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France
[2] LISN, Université Paris-Saclay, CNRS, Gif-sur-Yvette, France
[3] Université Paris Saclay, Gif-sur-Yvette, France
[4] LAMSADE, CNRS and Université Paris-Dauphine, PSL, Paris, France

alexandre.combeau@universite-paris-saclay.fr

Résumé

Dans un contexte où la nutrition est essentielle à la prévention des maladies chroniques, il est crucial d'adapter les recommandations alimentaires aux besoins individuels. Pour répondre à cette complexité, nous avons développé NutriKG*, un graphe de connaissances intégrant les données des études INCA2 et INCA3 avec une ontologie détaillée. Cette approche permet d'inférer de nouvelles informations et de combler les lacunes des données existantes. L'intégration de règles SWRL et de schémas SHACL assure la cohérence des recommandations et leur explicabilité. Ainsi, l'utilisation de* NutriKG *permettrait de faciliter la génération de recommandations alimentaires précises, personnalisées et scientifiquement fondées.*

Mots-clés

Graphes de Connaissances, Ontologies, Recommandation Alimentaire, Personnalisation, Explicabilité.

Abstract

In a context where nutrition is essential for preventing chronic diseases, it is crucial to tailor dietary recommendations to individual needs. To address this complexity, we developed NutriKG*, a knowledge graph integrating data from the INCA2 and INCA3 studies with a detailed ontology. This hybrid approach enables the inference of new information and helps bridge gaps in existing datasets. The integration of SWRL rules and SHACL schemas ensures the consistency and explainability of recommendations. Thus,* NutriKG *facilitates the generation of precise, personalized, and scientifically grounded dietary recommendations.*

Keywords

Knowledge Graphs, Ontologies, Food Recommendation, Personnalization, Explainability.

1 Introduction

Dans un contexte de préoccupations croissantes en matière de santé publique, notamment l'obésité, le diabète et les maladies cardiovasculaires, l'importance d'une alimentation saine est largement reconnue comme un facteur clé pour prévenir et atténuer ces problèmes [11]. Toutefois, les habitudes alimentaires sont influencées par de nombreux facteurs, notamment les préférences individuelles, les contraintes professionnelles, le niveau d'activité physique et les contextes culturels, rendant difficile l'adoption d'un régime alimentaire universellement efficace [3].

Face à cette diversité, il devient impératif de proposer des approches personnalisées fondées sur une analyse approfondie des habitudes alimentaires et des besoins spécifiques de chaque individu [12]. L'automatisation de cette tâche en combinant des méthodes d'apprentissage automatique et des méthodes issues du Web sémantique avec les graphes de connaissances offrirait une solution efficace et explicable [9]. Un système de recommandation alimentaire, exploitant ces méthodes et technologies, pourrait générer des suggestions diététiques adaptées aux profils nutritionnels, aux antécédents médicaux et aux préférences alimentaires des utilisateurs [15].

L'intégration de données issues de multiples sources telles que les journaux alimentaires, les dossiers médicaux et les choix de mode de vie permettrait d'améliorer la précision et la pertinence des recommandations [16]. En fournissant des conseils diététiques scientifiquement fondés et personnalisés, un tel système pourrait favoriser des habitudes alimentaires plus saines et contribuer à la réduction de l'impact des maladies liées à la nutrition sur les systèmes de santé [2].

L'objectif d'une telle approche est d'accompagner les individus vers des choix alimentaires éclairés, favorisant un meilleur état de santé général et une augmentation de la qualité de vie. À plus grande échelle, ces recommandations personnalisées pourraient réduire la prévalence des maladies chroniques et améliorer la productivité globale de la société [10].

Pour développer un tel système, nous nous sommes appuyés sur les données de référence de l'étude INCA (*Étude Individuelle Nationale des Consommations Alimentaires*),

mises à disposition par l'ANSES (*Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail*). Ce jeu de données constitue une ressource précieuse pour le développement d'outils tels que les systèmes de recommandation pour la nutrition. Bien que ces données soient bien structurées et de haute qualité, leur volume reste limité.

Dans [6], nous avons exploité les données INCA2 [1] afin d'entraîner un modèle capable de capturer les habitudes alimentaires des individus et de générer des recommandations de menus sous forme de séquences de plats. Dans un premier temps, nous avons présenté un modèle de recommandation inspiré d'une approche de type filtrage collaboratif. Cette première approche a montré ses atouts pour la recommandation d'un repas de type petit déjeuner, mais elle a révélé ses limites sur les repas plus complexes comme le déjeuner et le dîner. Nous avons présenté, ensuite, une seconde approche qui modélise le caractère séquentiel du déjeuner et du dîner et qui permet d'améliorer la performance de la recommandation sur ces repas. Cet approche repose sur une architecture de réseaux de neurones récurrents (*Recurrent Neural Networks*, RNN) pour l'apprentissage du modèle et la génération des séquences, en tenant compte de divers contextes (par exemple, les repas petit-déjeuner, déjeuner, dîner ou encore la tranche d'âge des individus). Ce travail a notamment démontré que l'intégration du contexte de consommation dans la génération d'une recommandation améliore considérablement sa pertinence.

Afin d'améliorer la précision du système de recommandation, de pallier la rareté des données et de garantir des recommandations explicables, nous avons développé *NutriKG*, un graphe de connaissances structuré en deux parties : (i) une composante conceptuelle, représentée par une ontologie, et (ii) une composante instancielle, construite à partir des données issues des deux études *INCA2* et *INCA3* [2] en 2006-2007 et en 2014-2015, respectivement.

Le développement du graphe de connaissances s'est déroulé en quatre étapes principales : (i) la création d'une première version de l'ontologie focalisée sur les consommations, leur structuration et leur composition ; (ii) l' extension de l'ontologie avec la modélisation des connaissances concernant les individus, leur contraintes et préférences alimentaires ; (iii) l'évolution de l'ontologie pour la prise en compte de l'organisation des données de la source INCA3 qui se différencient des données INCA2 par l'absence de séquentialité journalière des consommations et enfin (iv) la création des graphes de données conformes à l'ontologie et intégrant à la fois les données INCA2 et INCA3.

Ce graphe de connaissances est également enrichi par un ensemble de règles SWRL (Semantic Web Rule Language) [3] et de schéma de validation SHACL (Shapes Constraint Language) [4] pour compléter les informations manquantes (e.g., il manque l'information explicite sur le régime alimentaire pour 92% des individus dans INCA2), pour vérifier la conformité d'une recommandation par rapport aux contraintes de l'individu et enfin pour pouvoir fournir une explication intelligible pour les utilisateurs.

L'organisation de cet article est la suivante : la section 2 présente les travaux connexes, suivie par la section 3, qui introduit les préliminaires ainsi que les jeux de données issus de l'état de l'art. Les contributions principales sont détaillées dans la section 4. Ensuite, une première preuve de concept ainsi que les premiers résultats d'évaluation basés sur des questions de compétences sont présentés dans la section 5. Enfin, la section 6 conclut cet article et propose plusieurs perspectives pour de futurs travaux.

2 Travaux connexes

L'utilisation des ontologies dans le domaine alimentaire permet une modélisation formelle des connaissances relatives aux aliments, à leur composition, à leur transformation, ainsi qu'à leurs interactions avec la nutrition et la santé. Ces représentations sémantiques facilitent la structuration, le partage et l'interopérabilité des données. Plusieurs initiatives ont ainsi vu le jour pour formaliser et organiser ces connaissances sous la forme d'ontologies et de graphes de connaissances pour différents services et applications des domaines agroalimentaire et en santé.

Vocabulaires. Parmi ces initiatives, AGROVOC [13] constitue l'un des vocabulaires contrôlés les plus largement utilisés. Maintenu par la FAO (Organisation des Nations unies pour l'alimentation et l'agriculture), il couvre un large spectre de domaines, allant de l'alimentation et de la nutrition à l'agriculture, la pêche, la foresterie et l'environnement. Structuré sous forme de thésaurus hiérarchique, AGROVOC permet une organisation fine des concepts et favorise l'interopérabilité avec d'autres systèmes de gestion de l'information. Il est notamment utilisé pour l'indexation et la recherche d'informations dans le domaine agroalimentaire.

Ontologies des aliments. D'autres ontologies, plus spécifiques, se concentrent sur la représentation des aliments eux-mêmes et de leurs propriétés. FoodOn [4], intégrée à l'OBO Foundry [5], vise à décrire les entités alimentaires sous différents aspects : classification des aliments, composition nutritionnelle, propriétés physiques et rôle dans la chaîne alimentaire. Elle joue un rôle clé dans la structuration des données liées à la sécurité alimentaire, à la nutrition et à l'agriculture. FoodOn inclut également des catégories permettant de représenter les ingrédients, les produits transformés, ainsi que leurs relations avec les pratiques agricoles et les réglementations alimentaires. Cette approche facilite l'analyse des données et leur intégration dans des systèmes d'aide à la décision pour la gestion de la qualité et de la traçabilité des aliments.

Qualité et traçabilité. Outre la modélisation des aliments eux-mêmes, certaines ontologies ont été spécifiquement conçues pour l'évaluation des risques alimentaires. C'est notamment le cas de l'ontologie développée par l'ANSES

1. INCA2 : https://tinyurl.com/3zxjm6ca
2. INCA3 : https://tinyurl.com/487pj39d
3. SWRL : https://www.w3.org/submissions/SWRL/
4. SHACL : https://www.w3.org/TR/shacl/
5. https://obofoundry.org/

[8], qui structure les informations relatives aux expositions et aux dangers alimentaires. En s'appuyant sur l'alignement d'ontologies existantes, elle modélise des concepts tels que les contaminants alimentaires, les seuils de toxicité, les groupes à risque et les effets sanitaires potentiels. L'objectif est d'améliorer la précision des évaluations de risque et de permettre une automatisation partielle des analyses grâce aux raisonnements ontologiques.

Un autre axe de recherche concerne les ontologies dédiées aux processus de transformation des aliments. Certaines, comme celle développée par [7], modélisent les différentes étapes de préparation, cuisson et conservation des aliments. Ces ontologies intègrent des connaissances sur l'impact des procédés de transformation (e.g., fermentation, pasteurisation, surgélation) sur les qualités nutritionnelles et sanitaires des aliments. Elles décrivent également l'évolution des propriétés organoleptiques des aliments en fonction des traitements subis. Ces représentations sont particulièrement utiles pour les systèmes de traçabilité alimentaire et l'optimisation des procédés industriels, contribuant ainsi à une meilleure gestion de la qualité des produits transformés.

Dans le secteur de la production animale, l'ontologie développée dans le cadre du projet européen INTAQT [6] se focalise sur la viande de poulet, de boeuf et de produits laitiers. Elle définit et structure les concepts liés à la qualité, à la traçabilité et à la production de la viande. Elle intègre des données sur les races bovines, les méthodes d'élevage, les standards de qualité ainsi que les critères sensoriels tels que le goût, la texture et la couleur. Cette ontologie vise à améliorer l'interopérabilité des données entre les différents acteurs de la chaîne agroalimentaire et à renforcer la traçabilité des produits carnés et laitiers.

Ontologies et systèmes de recommandation en nutrition. Enfin, l'intégration des données alimentaires dans des systèmes de recommandation personnalisée constitue un défi majeur. FoodKG [5] est un graphe de connaissances conçu pour répondre à cet enjeu en combinant plusieurs sources de données alimentaires (i.e., ontologies, bases de données de recettes, publications scientifiques). Il permet d'exploiter des connaissances sémantiques pour affiner les recommandations alimentaires en fonction des préférences des utilisateurs, de la disponibilité des ingrédients et des contraintes nutritionnelles. Grâce à l'utilisation de SPARQL, FoodKG facilite l'interrogation et l'extraction d'informations pour générer des suggestions alimentaires plus pertinentes et adaptées aux consommateurs.

Ces différentes ontologies illustrent l'étendue des travaux menés pour structurer et modéliser les connaissances alimentaires. Elles constituent une base essentielle pour l'intégration et l'exploitation des données dans des systèmes d'information, des applications en nutrition et des services de recommandation alimentaire. Notre travail sur NutriKG est une extension de certaines de ces ressources existantes tel que FoodKG en fournissant un graphe de connaissance capturant davantage les comportements et les habitudes alimentaires des individus recueillis grâce aux initiatives INCA2 et INCA3.

6. https://www.sysaaf.fr/les-programmes-de-r-d/programmes-r-d-avi-en-cours/intaqt

3 Préliminaires

3.1 Ontologies et graphes de connaissances

Pour représenter sémantiquement les connaissances liées aux consommations nutritionnelles nous avons eu recours aux *ontologies* et aux *graphes de connaissances*. Une ontologie peut être définie comme une représentation formelle et structurée des connaissances d'un domaine, définissant des concepts, leurs relations et leurs propriétés, afin de permettre une compréhension partagée et une exploitation par des machines. On considère un graphe de connaissances comme une structure de données qui organise l'information sous forme de nœuds (entités) et d'arêtes (relations), facilitant l'intégration, le raisonnement et l'extraction de connaissances à partir de sources hétérogènes.

Dans la définition 3.1, nous donnons une définition formelle d'un graphe de connaissances et des éléments de l'ontologie qui permettent de le structurer.

Définition 3.1. (Graphe de connaissances RDF). Nous considérons un graphe de connaissances RDF défini par un couple $(\mathcal{O}, \mathcal{G})$, où :

– $\mathcal{O} = (\mathcal{C}, \mathcal{P})$ est une ontologie représentée en OWL2 [7] et composée d'un ensemble de classes $\mathcal{C}$ et de propriétés $\mathcal{P}$ pouvant être soit de type `owl:objectProperty`, dont le domaine et le co-domaine sont des classes, ou de type `owl:dataTypeProperty`, dont le domaine est une classe et le co-domaine est un type de données atomique (e.g, date, string, integer).

– $\mathcal{G}$ est un ensemble de faits représenté par des triplets de la forme {(sujet, propriété, objet) | sujet $\in \mathcal{I}$, propriété $\in \mathcal{P}$, objet $\in \mathcal{I} \cup \mathcal{L}$}, où $\mathcal{I}$ est l'ensemble d'instances de classes $c \in \mathcal{C}$ désignées par des IRI (Internationalized Resource Identifier), $\mathcal{P}$ est l'ensemble des propriétés, et $\mathcal{L}$ est l'ensemble des littéraux (tels que les nombres et les chaînes de caractères). On notera $I_C \subseteq \mathcal{I}$ l'ensemble d'instances de la classe $C \in \mathcal{C}$.

Les ontologies peuvent être enrichies par des règles, comme celles exprimées en SWRL, afin d'inférer de nouvelles connaissances à partir des faits existants. SWRL permet de définir des règles en logique du premier ordre pouvant exprimer le fait d'identifier qu'un aliment contenant un ingrédient allergène doit être évité par un individu. L'intégration de ces règles permet d'inférer de nouvelles connaissances, d'affiner les recommandations ou encore de détecter des incohérences.

3.2 Les jeux de données et référentiels

La construction du graphe de connaissances nutritionnelles repose sur l'intégration de plusieurs ensembles de données. Parmi ceux-ci, les principales sources actuellement utilisées incluent les sources de données INCA2 et INCA3,

7. OWL : https://www.w3.org/TR/owl2-overview/

alignées respectivement aux référentiels CIQUAL et FoodEx2 [8]. Ces sources de données fournissent des informations essentielles sur les habitudes alimentaires, la composition nutritionnelle des aliments et leur classification.
En complément, l'expertise humaine est mobilisée pour enrichir le graphe par des classes et des relations du domaine nécessitant une validation ou une interprétation spécifique. Dans la suite nous présentons ces différentes sources de données et référentiels ainsi que leur exploitation pour la construction et l'évolution du graphe de connaissances.

3.2.1 Le jeu de données INCA2

Le jeu de données INCA2 est issu d'une enquête nationale visant à analyser les comportements alimentaires des individus âgés de 3 à 79 ans vivant en France métropolitaine. Il repose sur un échantillon de 4 079 participants, répartis en deux groupes : les enfants (3 à 17 ans) et les adultes (18 à 79 ans). L'étude prend en compte divers facteurs démographiques et socio-économiques, tels que la région, le sexe, la taille du ménage et l'âge des individus.
Pour les enfants, l'enquête recueille des informations spécifiques sur les habitudes alimentaires en dehors du domicile ainsi que sur les préférences alimentaires (p. ex. : consommation de lait, de fruits). Chez les adultes, des données complémentaires sont intégrées, notamment sur l'activité physique et la consommation de certains produits comme la cigarette.
Concernant la consommation alimentaire, INCA2 s'appuie sur un journal alimentaire couvrant sept jours consécutifs, où chaque individu renseigne l'ensemble de ses repas : petit-déjeuner, déjeuner, dîner, ainsi que les collations intermédiaires. En complément, un questionnaire permet de collecter des informations détaillées sur les facteurs socio-économiques et les habitudes de vie.
Ce jeu de données comprend 1 280 références alimentaires et boissons, chaque élément étant associé à un vecteur nutritionnel extrait de la base CIQUAL 2008, qui fournit des données précises sur la composition nutritionnelle des aliments.

3.2.2 Le jeu de données INCA3

L'enquête INCA3, menée entre 2014 et 2015, repose sur un échantillon plus large de 5 855 participants, comprenant 2 698 enfants (0 à 17 ans) et 3 157 adultes (18 à 79 ans). Contrairement à INCA2, qui s'appuyait sur un suivi alimentaire continu sur sept jours, INCA3 adopte une méthodologie différente en collectant les informations sur deux à trois journées de 24 heures non consécutives. Cette approche, bien que plus souple, rend difficile la comparaison directe des résultats entre les deux enquêtes, comme le souligne la documentation officielle d'INCA3.
INCA3 apporte également une amélioration notable en intégrant des données plus détaillées sur divers aspects : facteurs socio-économiques, habitudes alimentaires, activité physique, état de santé et préférences alimentaires des enfants. L'enquête permet ainsi une analyse plus fine et approfondie des comportements alimentaires.
Grâce à sa richesse et à son niveau de détail, INCA3 a été utilisé dans un second temps pour enrichir la partie de notre ontologie dédiée aux consommateurs. Son exploitation permet d'améliorer la personnalisation des recommandations alimentaires en prenant en compte les besoins nutritionnels individuels ainsi que les facteurs liés au mode de vie.

3.2.3 Le jeu de données FoodEx2

Le jeu de données FoodEx2 est un système de classification des aliments développé par l'Autorité européenne de sécurité des aliments (EFSA). Conçu pour fournir un cadre standardisé de catégorisation et de codage des produits alimentaires, il facilite la collecte, l'analyse et l'interopérabilité des données dans les domaines de la sécurité alimentaire et de la nutrition.
Ce système repose sur une structure hiérarchique détaillée, attribuant à chaque produit alimentaire une description précise et un code spécifique. Cette organisation garantit une identification rigoureuse des aliments et assure la comparabilité des données entre différentes études et bases de données. Grâce à sa précision et à sa flexibilité, FoodEx2 est largement utilisé dans la recherche scientifique, la réglementation et la surveillance de la santé publique, contribuant ainsi à améliorer la fiabilité des informations nutritionnelles et alimentaires.
Dans le cadre de l'enquête INCA3, les aliments recensés sont déjà associés aux codes FoodEx2, permettant une classification détaillée. Les données incluent des informations sur le groupe, le sous-groupe, le sous-sous-groupe ainsi que le code spécifique FoodEx2, enrichi de facettes permettant de préciser davantage les caractéristiques des produits alimentaires.

3.2.4 Le référentiel CIQUAL

Le référentiel CIQUAL [9] est une base de données développée par l'ANSES, fournissant des informations sur la composition nutritionnelle des aliments consommés en France. Il répertorie plusieurs centaines d'aliments avec des données sur les macronutriments, minéraux, vitamines et autres composés. CIQUAL fournit des données sur la composition de plusieurs centaines d'aliments, couvrant un large éventail de catégories alimentaires (produits bruts, transformés, plats préparés, etc.). Chaque aliment est décrit par une série de paramètres nutritionnels, incluant :

- Macronutriments : Protéines, lipides, glucides, fibres alimentaires
- Éléments minéraux : Calcium, fer, magnésium, sodium, etc.
- Vitamines : Vitamine C, vitamines du groupe B, vitamine A, etc.
- Autres composés : Acides gras, sucres, additifs alimentaires

Les données proviennent d'analyses en laboratoire, de l'industrie agroalimentaire et de sources scientifiques. CIQUAL est mis à jour régulièrement et est utilisé en épidémiologie nutritionnelle, pour l'évaluation des apports alimentaires, le développement de produits et la modélisation

8. FoodEx2 :https://agroportal.lirmm.fr/ontologies/FOODEX2

9. https://ciqual.anses.fr/

des risques.

4 Le graphe de connaissances NutriKG

Dans cette section nous décrivons les éléments principaux qui composent le graphe de connaissances NutriKG ainsi que la méthodologie de sa construction. Nous présentons tout d'abord l'ontologie puis la méthodologie de construction de NutriKG.

4.1 L'ontologie NutriKG

Comme le montre la Figure 1, l'ontologie modélise deux aspects majeurs de la consommation alimentaire. La partie supérieure représente la composition des consommations, en mettant l'accent sur les aliments et les nutriments, tandis que la partie inférieure est dédiée à la modélisation des individus, intégrant leurs préférences, ainsi que leurs contraintes sanitaires et personnelles.

A) Consommations – classes et relations. Une consommation est modélisée à travers plusieurs classes interconnectées. La classe centrale, `FullDayConsumption`, représente l'ensemble des repas consommés par un individu au cours d'une journée (i.e., petit-déjeuner, déjeuner, collation et dîner). Cette classe est liée à `FoodComposition`, qui décrit pour chaque repas l'ensemble des plats consommés.

Les aliments consommés peuvent être simples, comme *eau* ou *banane*, ou composés, comme *tarte au citron*. Chaque aliment est relié à sa classification `foodex2`, permettant d'accéder à l'ensemble des informations fournies par FoodEx2, notamment la composition nutritionnelle et les différentes facettes associées à un plat (e.g., *mode de préparation, origine géographique*).

À chaque aliment est également associée sa composition nutritionnelle, obtenue par l'alignement des données INCA2 à CIQUAL. Les quantités des nutriments sont représentées sous forme de valeurs numériques accompagnées de leurs unités de mesure.

Un élément clé de la modélisation concerne la séquentialité des aliments consommés au sein d'un repas et des repas successifs au fil des jours. Cette dimension, issue des données INCA2 et INCA3, est intégrée dans l'ontologie à travers plusieurs propriétés temporelles et relationnelles. Celles-ci permettent de représenter à la fois l'ordre de consommation des aliments au sein d'un même repas et la chronologie des consommations journalières.

Parmi ces propriétés, la classe `FullDayConsumption` possède la propriété `before`, une propriété de type objet qui relie une instance de `FullDayConsumption` à une autre, correspondant à la consommation du jour précédent (qui ne correspond pas nécessairement au jour immédiatement antérieur).

En complément, les propriétés `hasBeginning` et `hasDuration` permettent de modéliser la séquence des aliments au sein d'un même repas ainsi que la durée associée à la consommation de chaque partie du repas (e.g., entrée, plat principal, boisson, dessert).

B) Individus – classes et relations. Dans cette seconde partie de l'ontologie, nous modélisons les caractéristiques des individus pouvant influencer la recommandation alimentaire.

Tout d'abord, la classe `Individu` regroupe des informations générales telles que la tranche d'âge, l'indice de masse corporelle (BMI), le genre ainsi que son profil alimentaire où l'on peut retrouver des informations sur son régime alimentaire (e.g., `VeganDiet, KetoDiet, LooseWeightDiet`)

Ensuite, deux classes permettent de représenter les préférences alimentaires de l'individu. La classe `FoodPreferences` modélise l'attrait d'une personne pour un aliment avec un booléen. En complément, la classe `FoodInterests` capture les préférences liées aux modes de préparation des repas (e.g., faits maison) ainsi que la propension de l'individu à découvrir de nouveaux aliments.

Les aspects médicaux et les contraintes de santé sont également pris en compte. La classe `MedicalInformation` regroupe des informations telles que le poids, la taille, la consommation de tabac et la volonté de perdre du poids. Quant à la classe `Restrictions`, elle modélise les éventuelles restrictions médicales ou allergies alimentaires de l'individu (e.g., gluten, fruits de mer).

Enfin, les habitudes et contraintes personnelles sont représentées par deux classes supplémentaires. La classe `PhysicalActivity` décrit les niveaux d'activité physique de l'individu, tandis que la classe `FoodProfile` formalise ses comportements alimentaires ainsi que ses objectifs liés à la gestion du poids.

4.2 NutriKG : méthodologie de construction

Nous avons construit l'ontologie NutriKG en suivant le 2ème scénario de la méthodologie de construction d'ontologie NeOn [14], c'est-à-dire celui qui consiste en la réutilisation et la réingénierie de ressources non ontologiques (NOR). La construction de l'ontologie en Turtle (TTL) a été réalisée en utilisant Chowlk [1], un outil permettant de convertir une modélisation UML en RDF. Cette approche assure une traduction fidèle du modèle conceptuel en un format exploitable pour le web sémantique.

Une première version de l'ontologie et du graphe de connaissances a été conçue à partir des descriptions de consommations issues du jeu de données INCA2. Par la suite, en s'appuyant sur les informations détaillées dans INCA3, tant sur les habitudes de consommation que sur les caractéristiques des individus, cette ontologie initialement centrée sur la consommation a été enrichie. De nouveaux concepts et relations ont été intégrés afin d'inclure une représentation plus complète des individus et de leurs profils associés, renforçant ainsi la capacité du graphe à modéliser les interactions entre les consommateurs et leurs choix alimentaires.

Afin de maximiser la quantité de données disponibles pour l'entrainement du modèle du système de recommandation, nous avons fixé comme objectif la conception d'une ontologie unificatrice capable d'intégrer de manière homogène

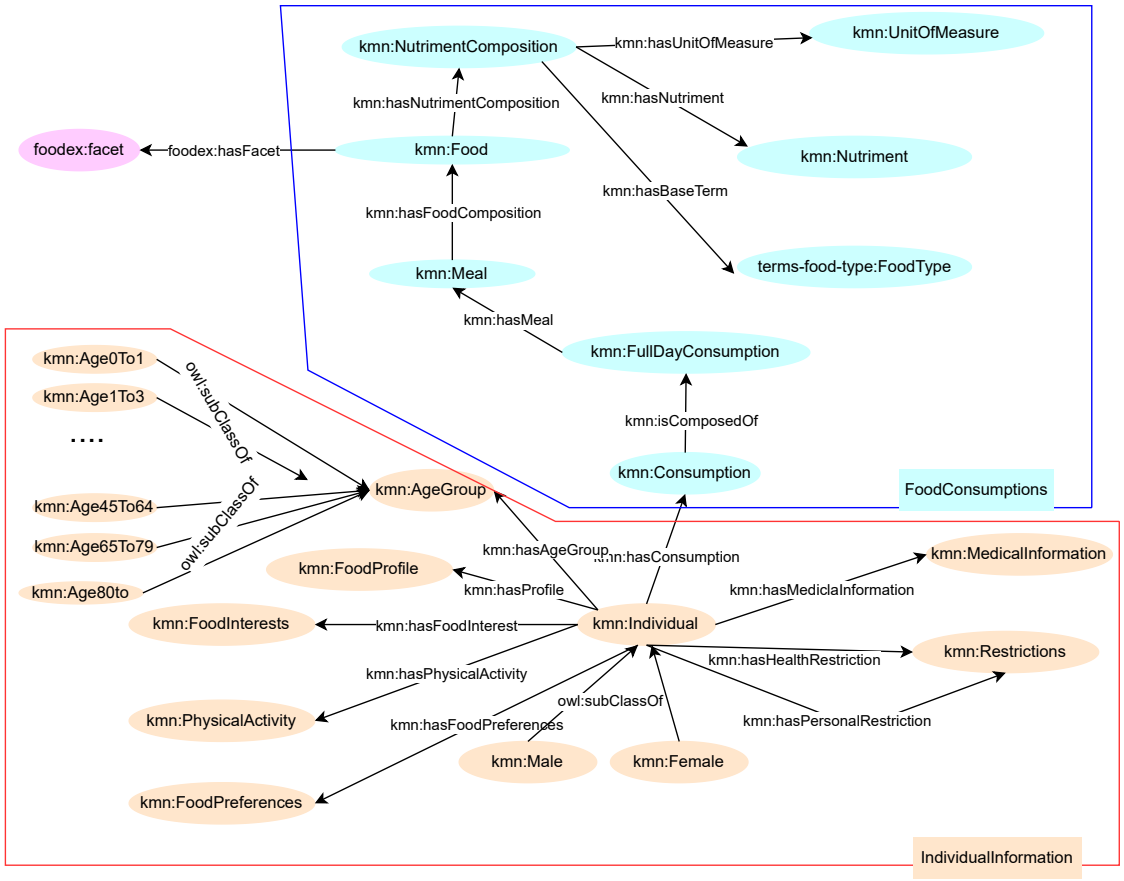

FIGURE 1 – L'ontologie du graphe de connaissances NutriKG issu des études INCA2 et de INCA3.

(sans perte d'informations) les jeux de données INCA2 et INCA3. Pour y parvenir, nous avons dû opter pour des choix de modélisation permettant de gérer l'hétérogénéité inhérente à ces sources de données. Cette hétérogénéité repose principalement sur deux aspects clés :

Période de collecte des données. Dans INCA2, la consommation alimentaire est enregistrée sur 7 jours consécutifs pour chaque individu, tandis que dans INCA3, seuls 3 jours non consécutifs sont pris en compte. Cette différence méthodologique a constitué un défi majeur dans la conception d'une ontologie capable de concilier ces deux approches et d'intégrer efficacement ces données.

Nomenclature et cartographie. Les bases INCA2 et INCA3 reposent sur des nomenclatures distinctes. Dans INCA3, la nomenclature a été associée à la classification FoodEx2. Ce dernier introduit des facettes, permettant d'affiner la description des aliments. Par conséquent, un même aliment dans INCA3 peut apparaître sous plusieurs déclinaisons, compliquant l'alignement avec INCA2. Le nombre exact d'instances présentes dans ces deux jeux de données est détaillé ci-dessous.

Dans le tableau 1 nous présentons quelques statistiques descriptives des deux jeux de données INCA2 et INCA3, i.e., nombre d'individus, nombre de jours, nombre de consommation et la taille d'une consommation.

Dataset	#Individus	#Jours	#Conso	Taille conso
INCA2	4079	7	80052	5.8
INCA3	3900	3	34964	5.7

TABLE 1 – Descriptions des données INCA

4.2.1 Homogénéisation des périodes de collecte des données

En raison des différences dans les méthodologies de collecte des données alimentaires entre INCA2 et INCA3, nous avons rencontré une incohérence structurelle dans la représentation des consommations. Pour y remédier, plusieurs ajustements ont été apportés à l'ontologie afin d'assurer une intégration plus cohérente et une meilleure exploitation des données.

La Figure 1 illustre la dernière version de l'ontologie mise à jour après résolution des incohérences. Dans la première version, un consommateur était associé à une consommation hebdomadaire, laquelle englobait des consommations journalières. Cette structuration a été revue afin de mieux correspondre à la nature des données collectées, notamment dans INCA3 où les consommations ne sont pas enregis-

trées sur une semaine complète. Ainsi, comme le montre la Figure 1, un consommateur est désormais associé à des consommations journalières indépendantes, sans contrainte temporelle spécifique à une semaine donnée.
Par ailleurs, pour mieux gérer l'historique des consommations, nous avons intégré l'ontologie Time (voir Figure 1). Cette amélioration est particulièrement pertinente pour INCA3, où les jours de consommation sont non consécutifs. Grâce à ce mécanisme, le système peut suivre les consommations passées et exploiter ces données pour formuler des recommandations alignées sur les habitudes alimentaires antérieures des utilisateurs, favorisant ainsi une alimentation plus équilibrée et personnalisée.
Enfin, nous avons restructuré la classification des aliments et des denrées afin d'adopter une hiérarchie conforme aux codes FoodEx2. Cette nouvelle organisation garantit une meilleure correspondance entre les différentes sources de données et facilite l'intégration des informations nutritionnelles dans le graphe de connaissances.

4.2.2 Homogénéisation des nomenclatures

Étant donné qu'INCA3 est déjà aligné sur FoodEx2, un standard largement adopté dans le domaine, nous avons choisi d'établir également une correspondance entre INCA2 et FoodEx2. Cependant, cette tâche a soulevé une difficulté majeure : les libellés des aliments dans INCA2 sont en français, tandis que ceux de FoodEx2 sont en anglais. Pour surmonter cet obstacle, nous avons exploré plusieurs options de traduction et constaté que l'API DeepL Translator [10] offrait les meilleurs résultats. Nous avons donc utilisé cette solution pour traduire les noms des aliments, des groupes alimentaires et des sous-groupes alimentaires d'INCA2 en anglais.
Afin d'automatiser la mise en correspondance entre INCA2 et FoodEx2, nous avons exploité l'application de codage intelligent FoodEx2. Cet outil, basé sur des réseaux neuronaux et des modèles Scapy [11], permet d'associer n'importe quel aliment à un code et à des facettes FoodEx2. Son code étant librement accessible sur GitHub, il constituait une solution adaptée à nos besoins. Cette approche a déjà été validée dans une étude précédente visant à établir une correspondance entre les produits alimentaires suédois et les codes FoodEx2.
Nous avons utilisé le modèle BaseTerm (BT) avec un seuil de similarité pour associer les aliments d'INCA2 aux codes FoodEx2. Les paramètres et les résultats détaillés de cette mise en correspondance sont présentés dans le tableau ci-dessous.
Dans le tableau 2 nous montrons les résultats des deux expériences menées. Dans la première, nous avons appliqué le modèle BT avec un seuil de 40 % (après plusieurs tests) en utilisant les libellés traduits des aliments (Libal), ce qui a permis de faire correspondre 1 196 instances de la nomenclature INCA2 sur un total de 1 343, laissant 147 instances non appariées. Afin d'améliorer ce résultat, nous avons exploité les informations sur les groupes et sous-groupes alimentaires dans une seconde expérience. Cette fois, nous avons utilisé le libellé du groupe alimentaire INCA2 (Libal_gr) avec le même modèle BT et le même seuil, ce qui nous a permis d'apparier 113 des 147 instances restantes.
L'objectif étant d'aligner les termes de base avec FoodEx2, cette approche s'est révélée efficace, ne laissant que 34 aliments non cartographiés, qui ont été traités manuellement à l'aide de connaissances expertes.

Carac. INCA2	Ins. mappées	Ins. Restantes
Libal	1196	147
Libal_gr	113	34

TABLE 2 – Alignement de INCA2 et Foodex2

5 Preuve de concept et évaluation

Dans cette section nous présentons nos résultats préliminaires de l'évaluation du graphe de connaissances NutriKG en nous appuyant sur des questions de compétences et sur trois applications possibles : (i) utilisation de schémas SHACL pour la vérification de la conformité des recommandations produites, (ii) l'utilisation du graphe et des règles SWRL pour inférer des informations manquantes, et (iii) production d'explications pour une recommandation de donnée.
L'ontologie de *NutriKG* a été mise à disposition de la communauté via *recherche.data.gouv.fr* et accessible rendu accessible avec le DOI https://doi.org/10.57745/037D7N. Les données du graphe de connaissances sont mises à disposition via un SPARQL endpoint généré par le triple store GraphDB [12].

5.1 Evaluation de l'ontologie et du graphe de connaissances

L'évaluation de l'ontologie est réalisée grâce à une série de questions de compétences issues des sites officiels de INCA 2 et de INCA 3 et un autre ensemble de questions fournies par les experts du domaine.

Q1	*Quelle quantité de nourriture mangent chaque jour les Français ?*
Q2	*Quel est le pourcentage d'obésité dans la population adulte ?*
Q3	*Quel est nombre d'individus déclarés végétariens ?*
Q4	*Quel est nombre d'individus allergiques aux oeufs ?*
Q5	*Quel est nombre d'individus n'ayant pas déclarés de régime alimentaire ?*

Pour la question Q2, nos résultats sont similaires aux observations de l'ANSES pour INCA2. Cependant, le prétraitement des données a eu un impact plus marqué sur INCA3, en particulier en raison du choix de ne conserver que les

10. https://www.deepl.com/en/pro-api
11. https://scapy.net/
12. https://graphdb.ontotext.com/

Question	Res. INCA2	Res. INCA3	évolution
Q1	2939g	2122g	817g
Q2	10.5%	16.7%	+6.2%
Q3	23	0	- 23
Q4	0	6	+ 6
Q5	3788	3435	- 353

TABLE 3 – Résultats des questions de compétences et expertes sur INCA2 et INCA3

Dataset	$\geq$ 18ans	15 - 17ans	11 - 14ans	3 - 10ans
INCA2	2939	1955	1898	1766
INCA3	2122	1775	1703	1415

TABLE 4 – Quantité de nourriture consommé (aliments et boissons) par jour en grammes selon l'âge sur les données INCA

repas du petit-déjeuner, du déjeuner et du dîner. Cette différence pourrait également refléter une évolution des habitudes de consommation.

5.2 Restrictions et profils

Cette ontologie peut également être intégrée à un système de recommandation, permettant ainsi d'évaluer la qualité d'une recommandation de repas pour un utilisateur en vérifiant le respect des contraintes qui lui sont associées.

Afin de modéliser les différentes contraintes alimentaires liées à la santé ou aux préférences personnelles des individus présentes dans les données INCA, nous les avons représentée par les classes `kmn:FoodProfile` et `kmn:Restrictions` (voir Figure 2).

Les restrictions définissent des contraintes négatives sur les aliments, influençant défavorablement leur recommandation à un utilisateur. Deux types de liens sont distingués : `kmn:hasHealthRestriction` et `kmn:hasPersonalRestrictions`, qui déterminent la sévérité de la restriction.

Une préférence personnelle peut, dans de rares cas, être prise en compte dans une recommandation, tandis qu'une contrainte de santé est strictement interdite. Par exemple, un plat contenant des œufs peut être recommandé à une personne ayant une restriction personnelle sur cet aliment, mais jamais si cette personne y est allergique.

La modélisation des habitudes et préférences de consommation s'appuie sur les profils `kmn:FoodProfiles`, qui traduisent une appétence particulière pour certains groupes d'aliments. Bien que ces profils soient déjà présents dans les données INCA3, aucune description ne précise les biais qu'ils devraient induire. Ainsi, notre objectif est de formuler un ensemble de contraintes afin de représenter au mieux les données réelles et l'état des connaissances expertes sur ces habitudes alimentaires.

Ces contraintes peuvent être utilisées pour contraindre la recommandation d'un individu. Les restrictions alimentaires sont exprimées à l'aide de schémas SHACL. Ci-dessous, nous présentons un exemple simplifié d'une restriction concernant les œufs. L'idée est de récupérer les consommations d'un utilisateur et de vérifier qu'aucune d'elles ne contient d'œufs. Pour distinguer une contrainte de santé d'une préférence personnelle, nous utilisons `sh:severity`, qui déclenche une violation de contrainte lorsqu'une contrainte de santé est enfreinte, et uniquement un avertissement dans le cas d'une préférence.

```
@prefix kmn: <http://example.org/kmn#> .
@prefix sh: <http://www.w3.org/ns/shacl#> .

kmn:NoEggRestrictionShape a sh:NodeShape ;
  sh:targetClass kmn:Individual ;
  sh:property [
    sh:path kmn:hasHealthRestriction ;
    sh:node kmn:NoEgg;] .
kmn:NoEggRestriction a sh:NodeShape ;
  sh:property [
    sh:path kmn:hasConsumption ;
    sh:node [
       sh:path kmn:hasFood ;
       sh:node [
         sh:path kmn:foodGroup ;
         sh:not [ sh:hasValue kmn:Egg ];] ;] ;
    sh:severity sh:Violation ;] .
```

Nous avons également enrichi le graphe de connaissances *NutriKG* avec un ensemble de règles en logique du premier ordre afin d'inférer et d'expliciter certaines connaissances sur les individus, auparavant implicites dans le graphe, telles que leur régime alimentaire. Ces règles, exprimées en SWRL, peuvent être exploitées par un raisonneur pour déduire de nouvelles connaissances.

Par exemple, un individu peut être défini comme végétarien s'il consomme une proportion de légumes supérieure à un seuil prédéterminé, ou à l'inverse, s'il consomme une quantité de produits d'origine animale inférieure à un seuil donné.

À titre d'exemple, et pour simplifier son expression, on peut formuler une règle en logique du premier ordre définissant les conditions à satisfaire pour qu'un individu soit considéré comme végétarien. Ainsi, pour un individu `ind` et ses consommations `foodC`, on définit une règle utilisant une fonction qui calcule la proportion de produits d'origine animale par rapport au nombre total de consommations, notée `countAnimal`. Si cette proportion dépasse un seuil `S`, on peut alors inférer le `FoodProfile` de type `VegetarianDiet`, que l'on associe à l'individu.

```
kmn:Individual(?ind) ^
kmn:hasProfile(?ind, ?profile)  ^
kmn:hasConsumption(?ind, ?foodC) ^
kmn:hasFood(?food, ?foodComp)^
kmn:foodGroup(?food, ?group)^
sqwrl:count(?totalCount, ?foodC)^
sqwrl:countDistinct(?aniCount,?foodC,
              ?group,?kmn:AnimalDerivedFood)^
swrlb:divide(?aniRatio,?aniCount,?totalCount)^
swrlb:greaterThanOrEqual(?aniRatio, S) =>
kmn:VegetarianDiet(?profile)
```

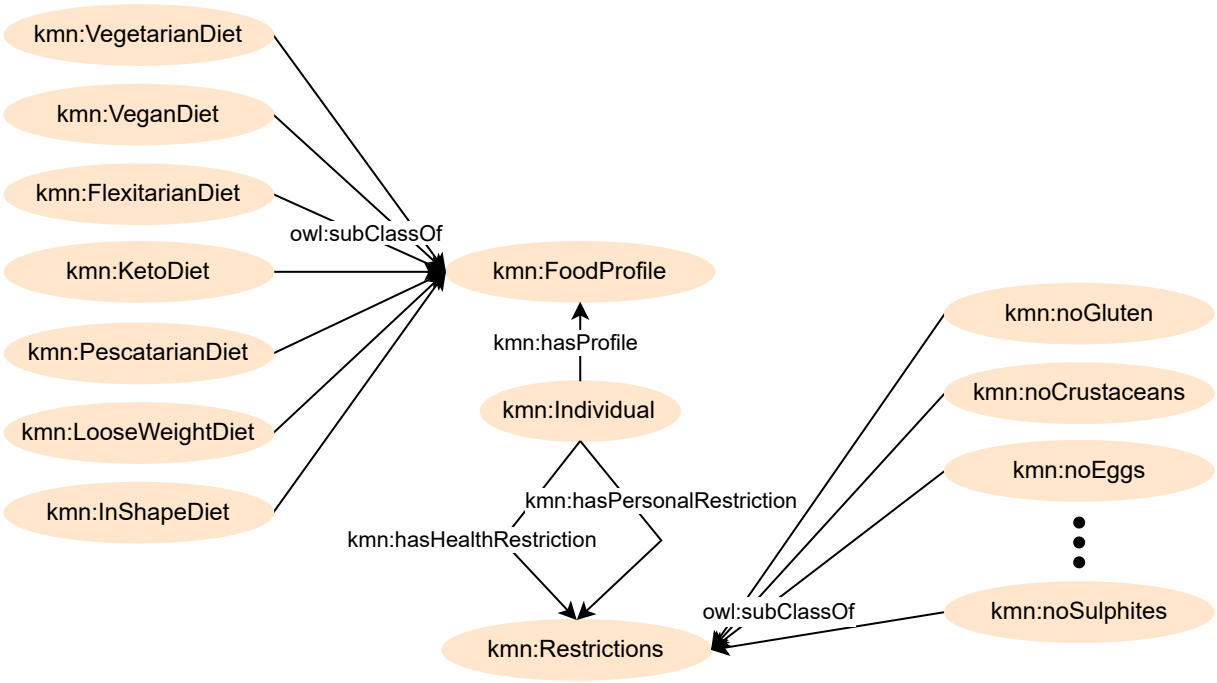

FIGURE 2 – NutriKg - Représentation des restrictions et profils alimentaires

À l'aide de ces profils, nous regroupons ensuite les aliments en fonction des profils des individus qui les consomment. Pour chaque aliment, nous calculons la proportion de profils qui l'incluent dans leur consommation. Cette analyse peut également être étendue à des classes et groupes d'aliments afin de faciliter la généralisation. Ce biais peut ensuite être exploité lors des recommandations, permettant ainsi au graphe d'enrichir les données préexistantes. Enfin, ces règles peuvent également servir de support pour la production d'explications pour les recommandations faites aux utilisateurs.

Au travers de cette première preuve de concept nous avons voulu montrer le potentiel de l'utilisation de la richesse du contenu de NutriKG dans le cadre de la recommandation alimentaire.

6 Conclusion et Travaux futurs

Dans un contexte où la nutrition est essentielle pour la prévention des maladies chroniques, la personnalisation des recommandations alimentaires est devenue un enjeu majeur. La diversité des facteurs influençant l'alimentation, tels que les préférences, les contraintes médicales et les habitudes culturelles, complexifie cette tâche. Dans cet article nous avons présenté NutriKG, un graphe de connaissances qui intègre des données de consommation issues des études INCA2 et INCA3, en les couplant à une ontologie, des règles SWRL et des schémas SHACL. Cette approche hybride associe l'organisation formelle des connaissances à la flexibilité des données réelles, ce qui permet non seulement d'inférer des informations manquantes, mais aussi de pallier les lacunes des jeux de données existants.

L'évaluation préliminaire de NutriKG montre la pertinence de la formalisation des connaissances et la structuration des données au travers plusieurs tâches : (i) réponses aux questions et en particulier à certaines questions de compétence de INCA2 et de INCA3 ; (ii) la capacité à inférer de nouvelles connaissances tels que les régimes alimentaires des individus, (iii) la vérification de conformité des recommandations vis-à-vis des restriction personnelles et médicales des individus grâce aux schémas SHACL et enfin (iv) à fournir des explications intelligibles pour des recommandations alimentaires.

Nous envisageons dans les travaux futurs plusieurs extensions de ce travail. Tout d'abord l'extension de l'alignement de l'ontologie NutriKG avec CIQUAL et d'autres ontologies existantes telles que FoodOn [4] ou encore celle de FoodKG[5]. L'intégration de NutriKG dans le système de recommandation tel que celui développé dans [6]. Cela peut être réalisé en post-traitement, d'abord, pour filtrer les recommandations incompatibles avec le profil/préférences de l'utilisateur, ensuite dans la phase d'apprentissage pour produire des recommandations plus pertinentes. Il serait également intéressant d'explorer la piste de l'utilisation du graphe de connaissances pour aider à l'augmentation de données. Enfin, nous souhaiterions proposer différentes définitions d'explications (e.g., contrastives, contrefactuelles) et des schémas de génération adaptés.

Remerciements

Cette recherche a été soutenue par l'Institut DATAIA Convergence dans le cadre du Programme d'Investissement d'Avenir (ANR-17-CONV-0003) opéré par l'Université Paris-Saclay et par la graduate school ISN (Informatique et Science du Numérique) de l'université Paris Saclay.

Références

[1] Serge Chávez-Feria, Raúl García-Castro, and María Poveda-Villalón. Chowlk : from uml-based ontology conceptualizations to OWL. In Paul Groth, Maria-Esther Vidal, Fabian M. Suchanek, Pedro A. Szekely, Pavan Kapanipathi, Catia Pesquita, Hala Skaf-Molli, and Minna Tamper, editors, *The Semantic Web - 19th International Conference, ESWC 2022, Hersonissos, Crete, Greece, May 29 - June 2, 2022, Proceedings*, volume 13261 of *Lecture Notes in Computer Science*, pages 338–352. Springer, 2022.

[2] GBD 2017 Diet Collaborators. Health effects of dietary risks in 195 countries, 1990–2017 : a systematic analysis for the global burden of disease study 2017. *The Lancet*, 393(10184) :1958–1972, 2019.

[3] Tom Deliens, Peter Clarys, Ilse De Bourdeaudhuij, and Benedicte Deforche. Determinants of eating behaviour in university students : a qualitative study using focus group discussions. *BMC Public Health*, 14(1) :53, 2014.

[4] H. Dooley, E. J. Griffiths, G. S. Gosal, P. L. Buttigieg, R. Hoehndorf, M. C. Lange, L. M. Schriml, F. S. L. Brinkman, and W. W. L. Hsiao. Foodon : a harmonized food ontology to increase global food traceability, quality control and data integration. *npj Science of Food*, 2 :23, 2018.

[5] Steven Haussmann, Oshani Seneviratne, Yu Chen, Yarden Ne'eman, James Codella, Ching-Hua Chen, Deborah L. McGuinness, and Mohammed J. Zaki. Foodkg : A semantics-driven knowledge graph for food recommendation. In *The Semantic Web – ISWC 2019 : 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II*, page 146–162, Berlin, Heidelberg, 2019. Springer-Verlag.

[6] Noémie Jacquet, Vincent Guigue, Cristina E. Manfredotti, Fatiha Saïs, Stéphane Dervaux, and Paolo Viappiani. Modélisation du caractère séquentiel des repas pour améliorer la performance d'un système de recommandation alimentaire. In *Extraction et Gestion des Connaissances, EGC 2024, Dijon, France, January 22-26, 2024*, volume E-40 of *RNTI*, pages 131–142. Editions RNTI, 2024.

[7] Moise Kombolo, Jérémy Yon, François Landrieu, Brigitte Richon, Sophie Aubin, and Jean-François Hocquette. Le Thésaurus de la viande : un nouvel outil accessible à tous Une nouvelle ressource sémantique répondant aux principes de la science ouverte : le thésaurus de la viande comme outil informatique de dialogue entre les acteurs de la filière. *Viandes et Produits Carnés*, March 2022.

[8] Myriam Merad, Sophie S. Allain, Jean-Christophe Augustin, Sandrine Blanchemanche, Gilles Bornert, Michel Federighi, Michel Gautier, Guillier Laurent, Nicole Hagen-Picard, Laïla Lakhal, Eric Marchioni, Régis Pouillot, and Brigitte Roudaut. Hiérarchisation des dangers biologiques et chimiques dans le but d'optimiser la sécurité sanitaire des aliments : Méthodologie et preuve de concept. Technical Report Saisine n°2016-SA-0153, Anses, 2020.

[9] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. A survey on food computing. *ACM Computing Surveys*, 52(5) :1–36, 2019.

[10] Carlos A. Monteiro, Geoffrey Cannon, Mark Lawrence, Maria Laura da Costa Louzada, and Priscila Pereira Machado. Ultra-processed foods, diet quality, and health using the nova classification system, 2019.

[11] Dariush Mozaffarian. Dietary and policy priorities for cardiovascular disease, diabetes, and obesity : a comprehensive review. *Circulation*, 133(2) :187–225, 2016.

[12] José M. Ordovás, Lynnette R. Ferguson, Esmond S. Tai, and John C. Mathers. Personalised nutrition and health. *BMJ*, 361 :bmj.k2173, 2018.

[13] Sachit Rajbhandari and Johannes Keizer. The agrovoc concept scheme – a walkthrough. *Journal of Integrative Agriculture*, 11(5) :694–699, 2012.

[14] Mari Carmen Suárez-Figueroa. Neon methodology for building ontology networks : specification, scheduling and reuse. In *DISKI*, 2011.

[15] Christoph Trattner and David Elsweiler. Food recommender systems : important contributions, challenges and future research directions. *arXiv preprint*, arXiv :1711.02760, 2017.

[16] Ying Zhu, Xiaodong Li, and Fei Wang. Personalized nutrition recommendation algorithm based on knowledge graph. *IEEE Access*, 8 :202778–202788, 2020.