TOTh 09

Terminologie & Ontologie : Théories et Applications

Actes de la troisième conférence TOTh - Annecy - 4 & 5 juin 2009

Institut Porphyre
Savoir et Connaissance

Dans un monde où la communication et le partage d'information sont au cœur de nos activités, les besoins en terminologie se font de plus en plus pressants. Il est devenu impératif d'identifier les termes employés et de les définir de façon consensuelle et cohérente tout en préservant la diversité langagière.

La terminologie, en tant que discipline scientifique, se fonde sur une conceptualisation d'un domaine et sur les mots pour en parler. Elle se doit donc de concilier un point de vue linguistique et un point de vue ontologique. Elle doit également, dans une société numérique où les connaissances constituent la principale richesse, pouvoir être opérationnalisée à des fins de traitement de l'information.

Les conférences TOTh se situent dans le prolongement des colloques annuels de la Société française de terminologie organisés en décembre à Paris (Ecole normale supérieure de la rue d'Ulm). Planifiées à mi-parcours, au mois de juin à Annecy (Polytech'Savoie), elles en complètent l'offre et proposent des conférences avec appel à communications, comité de lecture et publication des actes.

Les conférences TOTh ont pour objectif de rassembler industriels, chercheurs, utilisateurs et formateurs dont les préoccupations relèvent à la fois de la terminologie et de l'ontologie et, de façon plus générale, de la langue et de l'ingénierie des connaissances. Elles se veulent un lieu d'échange et de partage où sont exposés problèmes, solutions et retours d'expériences tant sur le plan théorique qu'applicatif ; ainsi que les nouvelles tendances et perspectives des disciplines associées : terminologie, langues de spécialité, linguistique, intelligence artificielle, systèmes d'information, ingénierie collaborative, etc.

Christophe Roche, Président du Comité Scientifique

http://www.porphyre.org

Institut Porphyre
Savoir et Connaissance

ISBN 978-2-9536168-0-4
EAN 9782953616804

Publications précédentes

TOTh 2007

Actes de la première conférence TOTh - Annecy - 1er juin 2007

TOTh 2008

Actes de la deuxième conférence TOTh - Annecy – 5 et 6 juin 2008

Commandes à adresser à : toth@porphyre.org

Titre : TOTh 2009. *Actes de la troisième conférence TOTh - Annecy – 4 & 5 juin 2009*
Editeur : Institut Porphyre, *Savoir et Connaissance*
http://www.porphyre.org
Annecy, 2009
ISBN 978-2-9536168-0-4
EAN 9782953616804

Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2009

Annecy – 4 & 5 juin 2009

avec le soutien de :

- Société française de terminologie
- Association Européene de Terminologie
- Ecole d'ingénieurs Polytech'Savoie – Université de Savoie
- Université de Sorbonne nouvelle
- Association EGC (Extraction et Gestion des Connaissances)
- ISKO (International Society for Knowledge Organization) France

Institut Porphyre
Savoir et Connaissance

http://www.porphyre.org

Comité scientifique

Président du Comité Scientifique : Christophe Roche

Comité de pilotage

Loïc Depecker	Professeur, Université de Sorbonne nouvelle
André Manificat	Directeur, GRETh
Christophe Roche	Professeur, Université de Savoie
Philippe Thoiron	Professeur émérite, Université de Lyon II

Comité de programme

Bruno de Bessé	Professeur, Université de Genève
Pierre Blanc	EDF SEPTEN
Danièle Bourcier	CNRS, CERSA Paris
Marc van Campenhoudt	Professeur, Termisti, ISTI, Bruxelles
Danielle Candel	CNRS, Université Paris Diderot
Stéphane Chaudiron	Professeur, Université de Lille III
Viviane Cohen	France Télécom, Paris
Rute Costa	Professeur, Université Nouvelle de Lisbonne
Luc Damas	MCF, Université de Savoie
Sylvie Després	MCF, Université Paris XIII
François Gaudin	Professeur, Université de Rouen
Anne-Marie Gendron	Chancellerie fédérale suisse, Section de terminologie
Jean-Yves Gresser	ancien Directeur à la Banque de France
Ollivier Haemmerlé	Professeur, Université de Toulouse
Jean-Paul Haton	Professeur, Université de Nancy 1
Michèle Hudon	Professeur, Université de Montréal
John Humbley	Professeur, Université Paris 7
Michel Ida	Directeur MINATEC, CEA
Hendrik Kockaert	Professeur, Lessius Hogeschool (Anvers)
Michel Léonard	Professeur, Université de Genève
Pierre Lerat	Professeur honoraire, Université Paris XIII
Widad Mustafa	Professeur, Université de Lille III
Henrik Nilsson	Terminologicentrum TNC, Suède
Jean Quirion	Professeur, Université du Québec en Outaouais
Renato Reinau	Suva, Lucerne
François Rousselot	MCF, Université de Strasbourg
Gérard Sabah	CNRS, Orsay
Michel Simonet	CNRS Grenoble
Marcus Spies	Professeur, Université de Munich
Dardo de Vecchi	Professeur associé, Euromed-Management

Comité d'organisation :

Responsable : Luc Damas

Samia Chouder, Joëlle Pellet

Avant propos

Dès la troisième édition, les conférences TOTh ont trouvé une structuration qui traduit bien à la fois le caractère scientifique et pluridisciplinaire de la terminologie et l'intérêt de notre communauté pour d'autres domaines partageant des préoccupations communes.

Ainsi, la conférence d'ouverture a été donnée par une personnalité invitée issue d'une discipline différente de la nôtre – ici la phylogénèse – mais pour laquelle le langage et la pensée jouent également un rôle primordial.

Les contributions se sont réparties naturellement, et par le jeu des évaluations de façon équitable, en trois groupes ayant donné lieu à trois sessions.

Le premier groupe a rassemblé les articles portant principalement sur la dimension linguistique de la terminologie. Ont été abordés l'extraction terminologique à partir de dictionnaire, la place accordée aux corpus dans la construction de terminologies, l'acquisition de connaissances à partir de textes et l'apport des ressources linguistiques issues du web.

La deuxième session s'est donc logiquement intéressée à la dimension conceptuelle de la terminologie. Les notions de concept, de relation, d'ontologie ont été au cœur des présentations portant sur les cartes conceptuelles pour les bibliothèques numériques, les relations dynamiques et les graphes conceptuels, l'alignement d'ontologies et l'accès multilingue aux ontologies.

Enfin, la troisième session a été consacrée à la présentation de plusieurs applications terminologiques pour des secteurs aussi différents que l'ingénierie nucléaire, l'informatique, le domaine bancaire ou l'agriculture biologique. Il est à souligner que ces applications ont permis d'aborder différents points théoriques tels que la variation terminologique, la diachronie ou la structure des dictionnaires.

La richesse des débats qui ont animé ces deux jours de conférence – chaque présentation, questions comprises, s'est vue allouer plus de quarante cinq minutes de temps de parole – a été certainement une des plus belles récompenses pour les participants de TOTh 2009.

Christophe Roche

Président du Comité Scientifique

Table des matières

SESSION 3

Construction et alignement d'ontologies pour évaluer le risque alimentaire

Liliana Ibanescu, Patrice Buche, Juliette Dibie-Barthélemy

Résumé : Nous présentons dans cet article un retour sur notre expérience de construction et d'alignement d'ontologies appliqué au domaine du risque alimentaire. Plus précisément, l'application concerne l'évaluation de l'exposition d'une population cible à un risque alimentaire réalisée avec le logiciel CARAT. Cette évaluation repose sur deux sources de données : une source de données de contamination chimique des aliments, appelée CONTA, et une source de données de consommation, appelée CONSO. Pour pouvoir calculer l'exposition d'une population cible à un risque alimentaire, les produits alimentaires de la source CONTA doivent être mis en correspondance avec les produits alimentaires de la source CONSO. Nous présentons dans ce papier le logiciel CARAT, ses sources de données, leurs référentiels, appelés ontologies, et le problème d'alignement de ces ontologies. Nous rappelons la méthode que nous avons proposée pour aligner ces ontologies, qui repose sur le modèle des graphes conceptuels. Nous présentons enfin les limites de cette méthode et proposons une nouvelle formalisation des ontologies pour les contourner.

Mots-clés : Représentation des connaissances, intégration de données, construction d'ontologie, alignement d'ontologies, risque alimentaire

1. Introduction

L'intégration des données permet d'accéder de manière unifiée à des sources multiples, hétérogènes en syntaxe, schéma ou sémantique. Le but de l'intégration de données est de faciliter l'accès et la réutilisation d'un ensemble de sources. La notion centrale sur laquelle reposent les recherches actuelles en intégration *sémantique* des données est la notion d'ontologie (Wache *et al.* 2001), (Ziegler *et al.* 2004), une ontologie étant un vocabulaire qui décrit un domaine d'intérêt et attribue un sens à ses termes (Gruber 1993). L'alignement des ontologies (Euzenat *et al.* 2007) est une solution pour résoudre l'hétérogénéité sémantique des données, en proposant des correspondances entre des entités sémantiques similaires de différentes ontologies.

Nous présentons dans cet article un retour sur notre expérience de construction et d'alignement d'ontologies dans le domaine du risque alimentaire. Plus précisément, nous présentons le logiciel CARAT (Buche *et al.* 2006) qui permet d'évaluer l'exposition d'une population cible à un risque alimentaire. Ce logiciel repose sur deux sources volumineuses de données : une source de données de contamination chimique des aliments, appelée CONTA, qui contient environ 2600 produits alimentaires différents, et, une source de données de consommations, appelée CONSO, qui contient environ 500 produits alimentaires. Dans CARAT, l'utilisateur est confronté au problème de l'intégration des deux sources de données qui reposent sur des référentiels distincts.

Nous présentons dans cet article l'architecture du logiciel CARAT, ses sources de données et leurs référentiels associés, appelés dans la suite ontologies, ainsi que le problème d'alignement de ses ontologies. Nous présentons ensuite notre méthode d'alignement des ontologies (Buche *et al.* 2008), qui repose sur le modèle des graphes conceptuels, et discutons des deux problèmes majeurs soulevés par cette méthode : i) un problème de modélisation : la description des produits alimentaires est au niveau instance (existentiel) ; ii) un problème concernant les résultats : les résultats ne sont pas satisfaisants et ils ne peuvent pas être comparés avec l'état de l'art. Enfin, nous proposons une nouvelle formalisation, en OWL (Dean *et al.* 2004), (Smith *et al.* 2004), des ontologies, et nous présentons la méthodologie utilisée pour construire ces ontologies à partir des sources de données existantes. Nous concluons et nous présentons les perspectives.

2. Le logiciel CARAT

Le logiciel CARAT (Buche *et al.* 2006) développé dans notre unité de recherche permet d'évaluer l'exposition des consommateurs à un risque alimentaire à partir de deux sources de données : une source de données de contamination chimique des aliments, appelée CONTA, et une source de données de consommation, appelée CONSO. Pour pouvoir calculer un niveau d'exposition, il faut définir trois éléments i) le contaminant (par exemple le méthyle-mercure); ii) le groupe d'individus étudié, appelé population cible (par exemple les enfants de moins de 15 ans); iii) le groupe de produits alimentaires (par exemple le groupe des poissons). Ensuite, il faut choisir la méthode de calcul statistique à utiliser pour évaluer le niveau de l'exposition de la population cible au contaminant étudié pour le groupe de produits alimentaires donné.

La Figure 7 présente l'architecture du système d'intégration de données du logiciel CARAT. Dans les deux sources de données, CONSO et CONTA, les noms des produits alimentaires ne sont pas identiques. Pour que CARAT puisse croiser les données de consommation avec les données de contamination, il est nécessaire de mettre en correspondance les noms de produits des référentiels de ces deux sources. Cette mise en correspondance est appelée alignement des ontologies CONSO et CONTA dans la Figure 7. Cet alignement permet ensuite d'interroger les deux sources de données en utilisant un seul vocabulaire. L'opération d'alignement repose sur la création de groupes de produits.

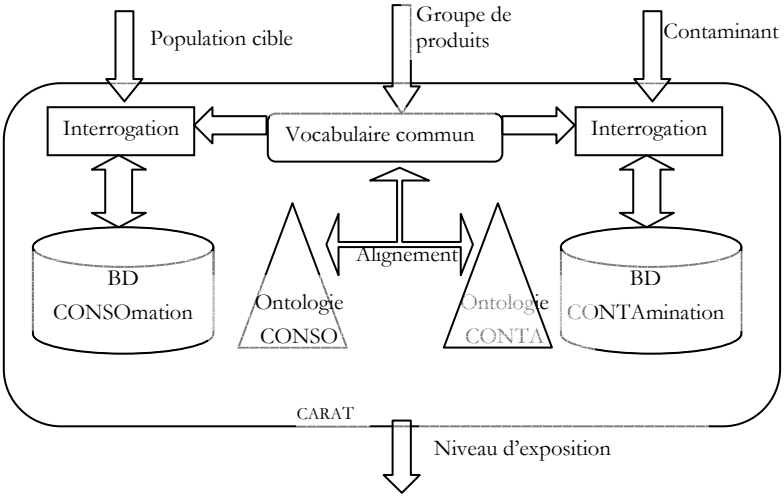

Figure 7. Architecture du système d'intégration de données du logiciel CARAT

La Figure 8 montre une capture d'écran du processus de construction d'un groupe d'aliments étudiés. Dans cette opération, un groupe de produits est défini comme deux ensembles de produits alimentaires : le premier ensemble contient les noms de produits du référentiel de la source CONSO, le second ceux du référentiel de la source CONTA. Cette opération permet donc d'aligner les référentiels de noms de produits des deux sources au niveau de granularité 'Groupe de produits'. Ces correspondances entre les noms de produits de la source CONSO et les noms de produits de la source CONTA sont actuellement établies à la main, par l'utilisateur du logiciel CARAT, ce qui représente un travail fastidieux. En effet, pour réaliser cette opération, l'utilisateur doit naviguer dans des référentiels volumineux : environ 500 termes pour la source CONSO et environ 2600 termes pour la source CONTA. Il est donc nécessaire de mettre à disposition de l'utilisateur un outil d'alignement semi-automatique des noms de produits alimentaires des deux sources afin de le soulager dans la tâche de création des groupes de produits.

Figure 8. Alignement entre les produits consommés et les produits contaminés pour définir le groupe de produits *Fruits et légumes secs*

Dans un premier temps, nous avons regardé s'il était possible d'effectuer une comparaison lexicale des noms de produits des deux ontologies CONSO et CONTA pour réaliser l'alignement. Le résultat de cette étude était loin d'être satisfaisant : l'alignement lexical des noms de produits a donné 13 correspondances exactes (égalité des chaînes de caractères) et 50

correspondances en comparant des sacs de mots (ensemble de mots lemmatisés contenus dans la chaîne de caractères correspondant au terme) sur un total de 3248 correspondances à trouver. Ce résultat médiocre s'explique principalement par la différence dans le niveau de granularité de la description des produits alimentaires dans les deux ontologies. L'ontologie CONSO (environ 500 termes désignant des noms de produits) est en effet beaucoup moins détaillée que l'ontologie CONTA (environ 2600 termes). Par exemple, *Poisson frais* dans l'ontologie CONSO doit être mis en correspondance avec *Cabillaud cru* dans l'ontologie CONTA. Nous avons alors décidé d'exploiter une description complémentaire des produits alimentaires disponible dans les deux sources de données pour effectuer l'alignement.

En effet, dans les deux sources de données, CONSO et CONTA, un produit alimentaire est décrit par une liste de triplets (produit, caractéristique, valeur). La Figure 9 présente des exemples de description. L'ensemble des triplets d'une source constitue l'ontologie des produits alimentaires associée à cette source.

Description d'un produit consommé **Poisson frais**	Description d'un produit contaminé **Cabillaud cru**
(Poisson frais, Présentation, Entier) (Poisson frais, Quel poisson ?, Cabillaud) (Poisson frais, Quel poisson ?, Saumon)	(Cabillaud cru, Origine de l'ingrédient principal, Morue ou cabillaud) (Cabillaud cru, Etat physique ou forme, Entier de forme naturelle) (Cabillaud cru, Méthode de conservation, Conservé par stockage en réfrigérant)

Figure 9. Exemples de triplets de description (produit, caractéristique, valeur) pour un produit consommé et respectivement un produit contaminé

Plus précisément, la source de données de consommation CONSO est produite par TNS Worldpanel France qui fournit un référentiel de 472 noms de produits alimentaires décrits avec 30 505 triplets (produit, caractéristique, valeur). Le nombre de caractéristiques distinctes est de 141 et celui de valeurs distinctes est de 15 116. Dans la source de données de contamination, CONTA, les produits alimentaires sont indexés selon le référentiel REGAL de l'AFSSA (Agence Française de Sécurité Sanitaire des Aliments) et décrits en utilisant Langual (Ireland *et al.* 2000), un thesaurus multilingue de description des aliments. Les 2595 noms de produits alimentaires sont décrits avec 25 069 triplets (produit, caractéristique, valeur). Le nombre des caractéristiques distinctes est de 14 et celui des valeurs distinctes est de 939. Toutes les valeurs d'une

caractéristique sont organisées dans une hiérarchie par la relation "sorte-de" (voir la Figure 10).

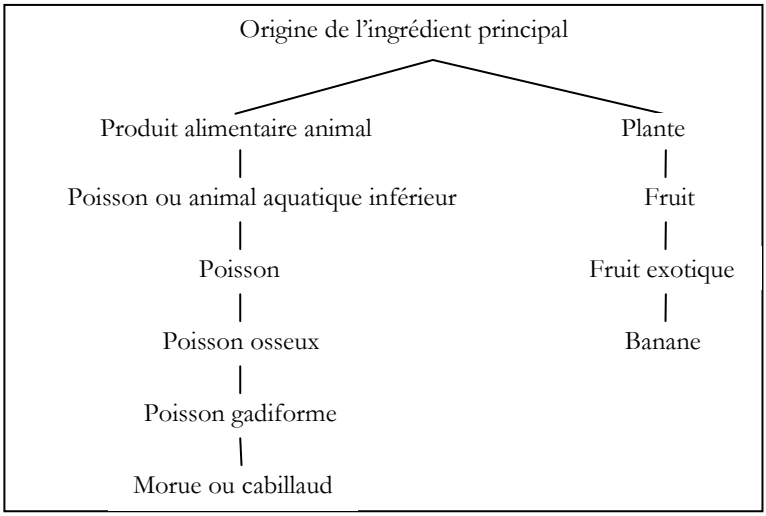

Figure 10. Un fragment de la hiérarchie des valeurs pour la caractéristique *Origine de l'ingrédient principal*

3. Alignement d'ontologies en utilisant les graphes conceptuels

Dans (Buche *et al.* 2008) nous proposons une méthode d'alignement d'ontologies, reposant sur le modèle des graphes conceptuels (Mugnier *et al.* 1996) dont nous rappelons dans cette section les grands principes.

Nous avons choisi le modèle des graphes conceptuels pour les raisons suivantes : i) les triplets (produit, caractéristique, valeur) de description d'un produit peuvent être aisément représentés sous la forme de graphes conceptuels; ii) la taxonomie des valeurs peut être directement représentée dans le support terminologique du modèle ; iii) la relation de projection du modèle peut être utilisée pour retrouver les alignements en exploitant les taxonomies de valeurs.

Les deux ontologies étudiées dans la méthode d'alignement ne sont pas symétriques. En effet, l'ontologie CONTA est stable dans le temps, elle est alors considérée comme l'ontologie cible, alors que l'ontologie CONSO évolue fréquemment et elle est considérée comme l'ontologie source. Le principe de la méthode proposée consiste à comparer la description d'un produit de l'ontologie CONSO à celle d'un produit de l'ontologie

CONTA en utilisant l'opération de projection du modèle des graphes conceptuels. Pour ce faire, les deux produits doivent être décrits dans le même vocabulaire. L'ontologie CONTA étant l'ontologie cible, nous avons choisi de traduire de manière semi-automatique chaque produit de l'ontologie CONSO dans le vocabulaire de l'ontologie CONTA. Plus précisément, la "traduction" d'un produit de l'ontologie CONSO est obtenue en remplaçant sa description en termes de caractéristiques et valeurs de CONSO par une description en termes de caractéristiques et valeurs similaires de l'ontologie CONTA. Les 4 grandes étapes de l'algorithme sont décrites ci-dessous.

La première étape consiste à trouver un alignement entre les caractéristiques des deux ontologies, en prenant en compte leurs valeurs. Par exemple : la caractéristique *Présentation* de l'ontologie CONSO est mise en correspondance avec la caractéristique *Etat physique ou forme* de l'ontologie CONTA, et la caractéristique *Quel poisson ?* est mise en correspondance avec la caractéristique *Origine de l'ingrédient principal*. L'établissement des correspondances entre caractéristiques est fait de manière semi-automatique à partir de similarités lexicales trouvées entre les valeurs qui leur sont associées. La validation manuelle de cette mise en correspondance par l'utilisateur est possible compte tenu du nombre restreint de caractéristiques. A la fin de cette étape, chaque caractéristique de l'ontologie CONSO est associée à une caractéristique de l'ontologie CONTA et chaque valeur de l'ontologie CONSO est associée à une liste de valeurs de l'ontologie CONTA pondérées par leur similarité lexicale (avec la valeur de l'ontologie CONSO).

Dans la deuxième étape, chaque description de produit de l'ontologie CONSO est "traduite" et représentée par un graphe conceptuel flou (Thomopoulos *et al.* 2003). Le support terminologique utilisé est défini de la manière suivante : i) l'ensemble des types de concepts contient les noms des produits des deux ontologies, l'ensemble des caractéristiques de l'ontologie CONTA, la hiérarchie des valeurs de l'ontologie CONTA et le type de concept *ValNum* (qui permet de représenter les valeurs numériques) ; ii) l'ensemble des types de relations contient les quatre types de relation *APourCarac*, *APourValeur*, *EstAnnoté* et *APourScore* ; iii) l'ensemble des marqueurs individuels contient les valeurs réelles. Le produit *Poisson frais* de l'ontologie CONSO (voir la Figure 9) est " traduit " et représenté par le graphe conceptuel de la Figure 11, avec les caractéristiques et les valeurs de l'ontologie CONTA. Dans cette traduction, chaque caractéristique de l'ontologie CONSO est remplacée par celle qui lui est associée dans l'ontologie CONTA et chaque valeur de

l'ontologie CONSO est remplacée par la liste pondérée de valeurs de CONTA associée.

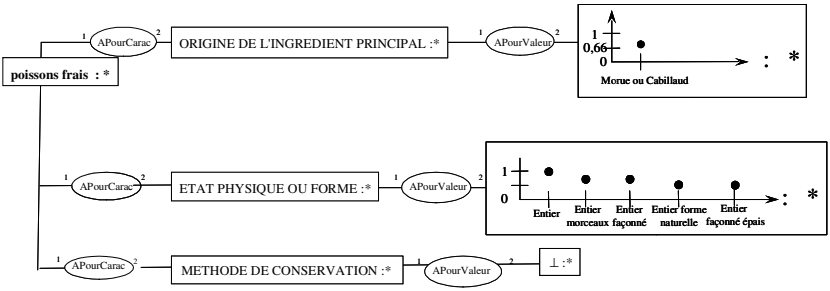

Figure 11. Représentation du produit *Poisson frais* de l'ontologie CONSO sous forme de graphe conceptuel

La troisième étape consiste à créer des règles pour représenter les produits de l'ontologie CONTA. A chaque produit de l'ontologie CONTA est associée une règle dont la partie condition contient la description du produit et dont la partie conclusion contient l'annotation qui est ajoutée à un graphe représentant un produit de l'ontologie CONSO si la règle est déclenchée. La règle de la Figure 12 est associée au produit *Cabillaud cru* de l'ontologie CONTA.

La quatrième étape consiste à appliquer toutes les règles créées dans l'étape précédente à l'ensemble des graphes associés aux produits de l'ontologie CONSO. L'application d'une règle a pour résultat d'ajouter au graphe associé à un produit de l'ontologie CONSO une annotation indiquant le produit de l'ontologie CONTA avec lequel il est aligné ainsi qu'un score d'alignement.. Par exemple, suite à l'application de la règle de la Figure 12, le produit *Poisson frais* de l'ontologie CONSO est annoté avec le produit *Cabillaud cru* de l'ontologie CONTA avec un degré de 0.5.

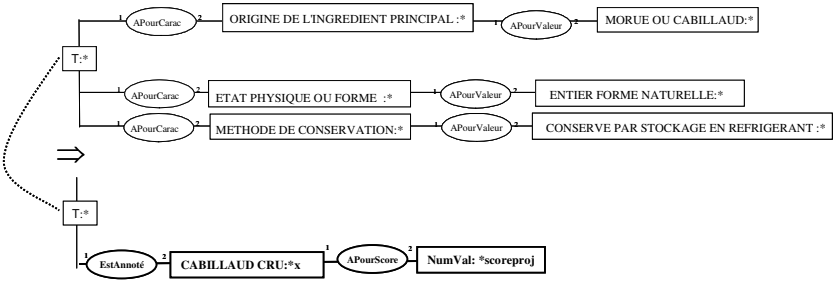

Figure 12. Règle associée au produit *Cabillaud cru* de l'ontologie CONTA

4. Limites de notre méthode d'alignement d'ontologies

Nous avons identifié deux problèmes majeurs soulevés par cette approche. Le premier problème concerne la modélisation. Si l'on met de côté l'extension floue proposée, la description sous forme de graphe conceptuel (voir la Figure 11) de la traduction d'un produit de l'ontologie CONSO, dans la formalisation proposée par (Mugnier *et al.* 1996), a une interprétation de fermeture existentielle conjonctive en logique des prédicats (il existe un poisson frais, …). Cela ne correspond pas à l'intuition que l'on peut avoir de la description d'un produit auquel on voudrait pouvoir associer une interprétation universelle (quel que soit le poisson frais, il a pour description…). Le deuxième problème concerne les résultats expérimentaux obtenus. Dans (Buche *et al.* 2008) nous avons calculé la précision[1] (2,08%) et la couverture[2] (77,04%) en comparant les résultats obtenus par notre méthode d'alignement à un alignement de référence que nous avons construit (manuellement) entre l'ontologie CONSO et l'ontologie CONTA. Des travaux ultérieurs, non encore publiés, nous ont permis d'améliorer la précision à 29,93%, avec une couverture de 59,85%, en ajoutant une étape supplémentaire de vérification de contraintes, pour éliminer des alignements incorrects. Ces résultats ne sont pas satisfaisants, dans la mesure où le travail de validation des résultats intermédiaires de l'alignement demandé à l'utilisateur est trop important. La raison essentielle que nous avons identifiée et qui pourrait expliquer ces mauvais résultats est l'existence de caractéristiques et de valeurs peu discriminantes mais génératrices de beaucoup de bruit. Par exemple, la caractéristique *Présentation* de l'ontologie CONSO qui, à la suite de la première étape de l'algorithme d'alignement, est mise en correspondance avec la caractéristique *Etat physique ou forme* de l'ontologie CONTA, permet par la suite de retrouver des alignements entre le produit *Poisson frais* de CONSO avec tous les produits de CONTA qui sont décrits avec le mot *Entier* parmi lesquels figurent tous les fromages. Cependant, pour pouvoir évaluer nos résultats expérimentaux, il faut pouvoir les comparer avec les résultats d'autres méthodes d'alignement. Or nos ontologies ne sont pas représentées dans un format adapté (RDF(S) ou OWL) pour pouvoir les aligner avec des outils existants d'alignement d'ontologies, comme ceux recensés dans (Euzenat *et al.* 2007), (Kalfoglou *et al.* 2005) et (Noy 2004).

1 La précision est le pourcentage d'alignements corrects générés par rapport à tous les alignements générés.

2 La couverture est le pourcentage d'alignements corrects générés par rapport à tous les alignements corrects.

Dans les deux sections suivantes nous présentons une nouvelle approche pour modéliser les ontologies CONSO et CONTA en OWL à partir des sources de données.

5. Construction de l'ontologie CONTA en OWL

Dans la source de données CONTA, qui est stockée sous la forme d'une base de données relationnelle, l'information concernant les produits alimentaires contaminés se trouve dans les trois tables suivantes (les clés primaires sont soulignées) :

produit_REGAL(nomProduit, nomFamille)

description(nomProduit, nomCaracteristique, valeurCaracteristique)

taxonomie_valeurs(valeur, valeurPere)

Les classes, les propriétés et les restrictions OWL associées à ces tables sont définies selon les étapes suivantes :

- Trois classes génériques, sous-classe de la classe Object, sont définies en OWL : *Famille*, *Produit*, *Valeur*.
- Pour chaque nom de famille 'nomFamille' distinct de la table produit_REGAL une nouvelle classe appelée 'nomFamille'[3] est créée. Cette classe est une sous-classe de la classe générique *Famille* et son étiquette est 'nomFamille'. Par exemple, pour la famille *Poissons et batraciens* le code généré en OWL est le suivant :

```
<owl:Class rdf:about= "#Poissons_et_batraciens">
   <rdfs:label xml:lang="fr"> Poissons et batraciens </rdfs:label>
   <rdfs:subClassOf rdf:resource="#Famille"/>
</owl:Class>
```

- Pour chaque nom de produit 'nomProduit' de la table produit_REGAL une nouvelle classe appelée 'nomProduit' est créée. Cette classe est une sous-classe de la classe générique *Produit* et son étiquette est 'nomProduit'. Par exemple, pour le produit *Cabillaud cru* le code généré en OWL est le suivant :

```
<owl:Class rdf:ID="Cabillaud_cru">
        <rdfs:label xml:lang="fr">Cabillaud, cru</rdfs:label>
        <rdfs:subClassOf rdf:resource="#Produit"/>
</owl:Class>
```

3 Aucun espace n'étant autorisé dans le nom des classes en OWL, on considère, dans la suite, que chaque espace dans le nom d'une classe est remplacé par le caractère spécial '_'.

- Pour chaque valeur 'valeur' d'une caractéristique décrite dans la table taxonomie_valeurs, une nouvelle classe de nom 'valeur' est créée. La représentation de la relation "sorte-de", codée par les tuples (valeur, valeurPere) de la table taxonomie_valeurs, est traduite en OWL par le fait que la classe 'valeur' est une sous-classe de la classe 'valeurPere'. Par exemple, le tuple (*Morue ou cabillaud*, *Poisson gadiforme*) de la table taxonomie_valeurs (qui code la hiérarchie de la Figure 4) et représenté en OWL par le code suivant :

```
<owl:Class rdf:about="#morue_ou_cabillaud">
   <rdfs:label xml:lang="fr">MORUE OU CABILLAUD</rdfs:label>
   <rdfs:subClassOf rdf:resource="# poisson_gadiforme"/>
</owl:Class>
```

- Les valeurs de chaque caractéristique sont organisées dans une hiérarchie dont la racine est une sous-classe de la classe générique *Valeur*. Par exemple, pour la hiérarchie de la caractéristique *Origine de l'ingrédient principal* (cf. Figure 4) le code généré en OWL est le suivant :

```
<owl:Class rdf:about="#Val_origine_ingredient_principal">
  <rdfs:label xml:lang="fr">Origine de l'ingrédient
principal</rdfs:label>
   <rdfs:subClassOf rdf:resource="#Valeur"/>
</owl:Class>
```

- La relation exprimée par la table produit_REGAL entre un nom de produit et un nom de famille est représentée en OWL par une propriété 'Appartient_a_famille' dont le domaine est la classe générique *Produit* et le co-domaine est la classe générique *Famille*.

```
<owl:ObjectProperty rdf:about="#APPARTIENT_A_FAMILLE">
<rdfs:domain rdf:resource="#Produit"/>
<rdfs:range rdf:resource="#Famille"/>
</owl:ObjectProperty>
```

- Chaque tuple (nomProduit, nomFamille) de la table produit_REGAL est représentée en OWL par une restriction sur la propriété APPARTIENT_A_FAMILLE, cette restriction étant ajoutée à la définition de la classe 'nomProduit'. Par exemple, le tuple (*Cabillaud cru*, *Poissons et batraciens*) est décrit dans la définition de la classe associée au produit *Cabillaud cru* par le code suivant en OWL :

```
<owl:Class rdf:ID="Cabillaud_cru">
```
… (cf. code de l'étape 3)

```
<rdfs:subClassOf>  <owl:Restriction>
 <owl:allValuesFrom rdf:resource="#Poissons_et_batraciens"/>
   <owl:onProperty>
     <owl:ObjectProperty rdf:about="#APPARTIENT_A_FAMILLE"/>
   </owl:onProperty>
</owl:Restriction> </rdfs:subClassOf>
</owl:Class>
```

- Pour chaque relation distincte de la table description entre un nom de produit 'nomProduit' et une caractéristique 'nomCaracteristique', une propriété appelée 'nomCaracteristique' est créée en OWL. Cette propriété a pour domaine la classe générique *Produit* et pour co-domaine la classe associée à la racine de la hiérarchie de valeurs de la caractéristique 'nomCaracteristique' (définie dans l'étape 5).

```
<owl:ObjectProperty
rdf:about="#ORIGINE_INGREDIENT_PRINCIPAL">
   <rdfs:domain rdf:resource="#Produit"/>
   <rdfs:range rdf:resource="# Val_origine_ingredient_principal "/>
</owl:ObjectProperty>
```

- Chaque tuple (nomProduit, nomCaracteristique, valeurCaracteristique) de la table description est représenté en OWL par une restriction sur la propriété 'nomCaracteristique', cette restriction étant ajoutée à la définition de la classe 'nomProduit'. Par exemple, le tuple (*Cabillaud cru*, *Origine de l'ingrédient principal*, *Morue ou cabillaud*) de la Figure 3 est décrit dans la définition de la classe associée au produit *Cabillaud cru* par le code suivant en OWL :

```
<owl:Class rdf:ID="Cabillaud_cru">
… (cf. code des étapes 3 et 7)
<rdfs:subClassOf>  <owl:Restriction>
   <owl:onProperty>
      <owl:ObjectProperty
rdf:about="#ORIGINE_INGREDIENT_PRINCIPAL "/>
   </owl:onProperty>
   <owl:allValuesFrom>
     <owl:Class rdf:ID="morue_ou_cabillaud">
   </owl:allValuesFrom>
</owl:Restriction> </rdfs:subClassOf>
</owl:Class>
```

L'ontologie CONTA générée en OWL compte 7 495 classes (dont 43 pour les familles, 2 595 pour les produits et 4 854 pour les valeurs) et 25 112 propriétés.

6. Construction de l'ontologie CONSO en OWL

Dans la source de données CONSO, qui est stockée sous la forme d'une base de données relationnelle, l'information concernant les produits alimentaires consommés se trouve dans les deux tables suivantes :

produit_SECODIP(nomProduit, nomFamille)

description(nomProduit, nomCaracteristique, valeurCaracteristique)

Les étapes de construction de l'ontologie CONSO sont à peu près les mêmes que pour l'ontologie CONTA. On peut identifier deux grandes différences : i) dans l'étape 4, la génération de la taxonomie de valeurs ne sera pas possible car les valeurs de l'ontologie CONSO ne sont pas organisées en hiérarchies ; ii) dans l'étape 9, la restriction des propriétés ne peut pas se faire avec l'opérateur **allValuesFrom,** mais avec l'opérateur **someValueFrom**, car, pour un produit donné, l'ensemble des tuples (nomProduit, nomCaracteristique, valeurCaracteristique) de la table description de l'ontologie CONSO ne donne pas une description d'un produit alimentaire, mais toutes les possibilités pour le décrire. Par exemple, pour le produit *Poisson frais* de la Figure 3 le code OWL généré est le suivant :

```
<owl:Class rdf:ID="POISSON_FRAIS">
<rdfs:label xml:lang="fr">POISSON FRAIS</rdfs:label>
<rdfs:subClassOf> <owl:Restriction>
        <owl:onProperty>
  <owl:ObjectProperty rdf:about="#PRESENTATION"/>
        </owl:onProperty>
        <owl:someValuesFrom rdf:resource="#entier"/>
</owl:Restriction> </rdfs:subClassOf>
<rdfs:subClassOf> <owl:Restriction>
        <owl:onProperty>
  <owl:ObjectProperty rdf:about="#QUEL_POISSON"/>
        </owl:onProperty>
        <owl:someValuesFrom rdf:resource="#cabillaud"/>
</owl:Restriction> </rdfs:subClassOf>
<rdfs:subClassOf> <owl:Restriction>
        <owl:onProperty>
          <owl:ObjectProperty rdf:about="#QUEL_POISSON"/>
        </owl:onProperty>
        <owl:someValuesFrom rdf:resource="#saumon"/>
</owl:Restriction> </rdfs:subClassOf>
</owl:Class>
```

L'ontologie CONSO générée en OWL compte 15 655 classes (dont 64 pour les familles, 472 pour les produits et 15 116 pour les valeurs) et 30 569 propriétés.

Cette représentation en OWL résout le problème de modélisation rencontré avec le modèle des graphes conceptuels dans la section précédente. L'interprétation logique de la description d'un produit dans la modélisation OWL proposée est bien de type universelle (quel que soit le cabillaud cru, il a pour description…).

7. Conclusion et perspectives

Dans le domaine du risque alimentaire, l'évaluation de l'exposition d'une population à un risque chimique requiert le croisement de sources de données de consommation et de contamination des aliments. Afin d'effectuer ce croisement, il est préalablement nécessaire d'aligner les noms de produits alimentaire utilisés dans les deux sources pour indexer les données. Nous avons présenté dans cet article un retour sur notre expérience de construction et d'alignement d'ontologies de produits alimentaires que nous aimerions utiliser pour étendre le système d'intégration de données du logiciel CARAT (Buche *et al.* 2006).

Pour trouver de manière semi-automatique les correspondances entre les noms des produits des deux sources nous avons proposé dans (Buche *et al.* 2008) une méthode d'alignement d'ontologies basée sur le modèle des graphes conceptuels (Mugnier *et al.* 1996). Cette méthode combine des techniques syntaxiques (comme la lemmatisation des noms de produits) avec des techniques structurelles utilisant la taxonomie des valeurs. Dans cet article nous avons rappelé les étapes de l'algorithme et nous avons mis en évidence et discuté des problèmes rencontrés : i) un problème de modélisation dû au choix initial du modèle des graphes conceptuels ; ii) un problème concernant les résultats expérimentaux qui ne sont pas satisfaisants.

Nous avons ensuite proposé dans cet article une nouvelle formalisation, en OWL (Dean *et al.* 2004) (Smith *et al.* 2004), des ontologies associées aux sources de données, en précisant les règles de transformation des métadonnées et données extraits de chaque base de données relationnelle en classes et propriétés. Notre méthode est automatique. Elle se différencie des propositions récentes dans le domaine de la génération automatique d'ontologies à partir d'une base relationnelle (Astrova 2007), (Lubyte *et al.* 2007). En effet, nous voulons conserver dans l'ontologie la représentation des taxonomies stockées dans les tables de la base en

utilisant la relation de spécialisation entre classes. Or ces méthodes représentent sous forme d'instances les tuples de données stockées dans les tables.

Les perspectives à court terme consistent à comparer les performances de notre méthode d'alignement avec celles de l'état de l'art grâce à la modélisation OWL de nos ontologies proposée dans cet article. Les perspectives à plus long terme sont d'étudier comment il est possible d'intégrer dans les méthodes d'alignement d'ontologies les évolutions périodiques de l'ontologie CONSO.

Bibliographie

Astrova I. (2007) : *Rules for Mapping SQL Relational Databases to OWL Ontologies*. *MTSR 2007*: 415-424

Buche P., Soler L., Tressou J. (2006) : *Le logiciel CARAT*. Dans : Bertail P., Feinberg M., Tressou J., Verger P., Analyse des Risques alimentaires, Lavoisier Tech&Doc, pp. 305-333

Buche P., Dibie-Barthélemy J., Ibanescu L. (2008) : *Ontology Mapping Using Fuzzy Conceptual Graphs and Rules*. ICCS'08 Supplement, pp. 17-24

Dean M., Schreiber G. (Editors) (2004) : *OWL Web Ontology Language Reference*, W3C Recommendation, 10 February 2004

Euzenat, J., Shvaiko, P. (2007) : *Ontology Matching*. Berlin: Springer

Gruber T.R. (1993) : *A Translation Approach to Portable Ontology Specifications*. Knowledge Acquisition, 5(2):199-220

Ireland, J. D., Moller, A. (2000) : *Review of International Food Classification and Description. Journal of Food Composition and Analysis*, 33, pp. 529-538

Kalfoglou, Y., Schorlemmer M. (2005) : *Ontology Mapping: The State of the Art*. Semantic Interoperability and Integration

Lubyte L., Tessaris S. (2007) : *Extracting Ontologies from Relational Databases*. Description Logics 2007

Mugnier, M., Chein, M. (1996) : *Représenter des connaissances et raisonner avec des graphes*. Revue d'Intelligence Artificielle 10 (1), 7-56

Noy, N. F. (2004) : *Semantic Integration: A Survey of Ontology-Based Approaches*. ACM SIGMOD Record, 33(4), pp. 65-70

Smith M. K., Welty C., McGuinness D.L . (Editors) (2004) : *OWL Web Ontology Language Guide*, W3C Recommendation, 10 February 2004

Thomopoulos T., Buche P., Haemmerlé O. (2003) : *Different Kinds of Comparisons between Fuzzy Conceptual Graphs*. ICCS 2003: 54-68

Wache H., Vögele T., Visser U., Stuckenschmidt H., Schuster G., Neumann H., Hübner S. (2001) : *Ontology-Based Integration of Information - A Survey of Existing Approaches*, pp. 108-117

Ziegler P., Dittrich K. R. (2004) : *Three Decades of Data Integration - All Problems Solved*, WCC 2004, 12, pp. 3-12

A propos des auteurs

Liliana Ibanescu
Mét@risk – INRA, UFR Informatique, AgroParisTech
16 rue Claude Bernard, F-75231 Paris Cedex 05
Liliana.Ibanescu@agroparistech.fr

Patrice Buche
Mét@risk – INRA
16 rue Claude Bernard, F-75231 Paris Cedex 05
Patrice.Buche@paris.inra.fr

Juliette Dibie-Barthélemy
Mét@risk – INRA, UFR Informatique, AgroParisTech
16 rue Claude Bernard, F-75231 Paris Cedex 05
Juliette.Dibie@agroparistech.fr