Actes des 25es journées francophones d'Ingénierie des Connaissances (IC 2014)

Catherine Faron Zucker

▶ **To cite this version:**

Catherine Faron Zucker (Dir.). Actes des 25es journées francophones d'Ingénierie des Connaissances (IC 2014). Catherine Faron-Zucker. Conférence IC, pp.300, 2014. hal-01010292

HAL Id: hal-01010292
https://inria.hal.science/hal-01010292

Submitted on 19 Jun 2014

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.

Actes des 25es journées francophones d'Ingénierie des Connaissances IC 2014

Catherine Faron Zucker

Présidente du comité de programme

Catherine Roussey

Présidente du comité d'organisation

Irstea & Université Blaise Pascal

Campus des Cézeaux

63170 Aubière, France

12-16 mai 2014

Avec le soutien scientifique et financier de :

Comité de programme

Présidente du comité de programme :
Catherine Faron Zucker, Maître de conférences à l'Université Nice Sophia Antipolis – Laboratoire I3S

Marie-Helene Abel, Université de Technologie de Compiègne - Laboratoire Heudiasyc,
Yamine Ait Ameur, École Nationale Supérieure de Mécanique et d'Aérotechnique - Laboratoire LISI
Patrick Albert, IBM France
Florence Amardeilh, Mondeca
Manuel Atencia, Université de Grenoble 1
Marie-Aude Aufaure, Ecole Centrale Paris - Laboratoire MAS
Nathalie Aussenac-Gilles, CNRS - Laboratoire IRIT
Bruno Bachimont, Université de Technologie de Compiègne - Laboratoire Heudiasyc
Jean-Paul Barthès, Université de Technologie de Compiègne - Laboratoire Heudiasyc
Aurélien Bénel, Université de Technologie de Troyes - Laboratoire Tech-CICO
NacéraBennacer, SUPELEC - Equipe E3S
Bertrand Braunschweig, INRIA Rennes-Bretagne Atlantique
Nathalie Bricon-Souf, Université Lille 2 - Centre d'Etudes et de Recherche en Informatique Médicale
Patrice Buche, INRA
Elena Cabrio, INRIA Sophia Antipolis
Jean-Pierre Cahier, Université de Technologie de Troyes - Laboratoire Tech-CICO
Sylvie Calabretto, Institut National des Sciences Appliquées de Lyon - Laboratoire LIRIS
Pierre-Antoine Champin, Université Claude Bernard Lyon 1 - Laboratoire LIRIS
Jean Charlet, Université Pierre et Marie Curie - AP-HP/INSERM UMR_S 872
Olivier Corby, INRIA Sophia Antipolis
Amélie Cordier, Université Claude Bernard Lyon 1 - Laboratoire LIRIS
Michel Crampes, Ecole des Mines d'Ales - Laboratoire LGI2P
Philippe Cudré-Mauroux, Université de Fribourg
Olivier Curé, Université de Marne-la-Vallée - Laboratoire d'informatique Gaspard-Monge
Célia Da Costa Pereira, Université Nice Sophia Antipolis - Laboratoire I3S
Mathieu D'Aquin, The Open University - Laboratoire KMi
Jérôme David, Université Pierre-Mendès-France - Laboratoire LIG
Sylvie Despres, Université Paris 13 - Laboratoire LIM&BIO
Rim Djedidi, Université Paris 13 - Laboratoire LIM&BIO
Jérôme Euzenat, INRIA Grenoble
Gilles Falquet, Université de Genève - Laboratoire ISI
Catherine Faron Zucker, Université Nice Sophia Antipolis - Laboratoire I3S
Cécile Favre, Université Lyon 2 - Laboratoire ERIC
Béatrice Fuchs, Université Jean Moulin Lyon 3 - Laboratoire LIRIS
Frédéric Fürst, Université de Picardie Jules Verne - Laboratoire MIS
Jean-Gabriel Ganascia, Université Pierre et Marie Curie - Laboratoire LIP6
Fabien Gandon, INRIA Sophia Antipolis
Aldo Gangemi, Université Paris 13 - Laboratoire LIPN et ISTC-CNR, Rome, Italie
Catherine Garbay, CNRS - Laboratoire LIG
Serge Garlatti, Telecom Bretagne
Alain Giboin, INRIA Sophia Antipolis
Monique Grandbastien, Université de Lorraine, Laboratoire LORIA

Christophe Guéret, Data Archiving and Networked Services
Ollivier Haemmerlé, Université de Toulouse Le Mirail - Laboratoire IRIT
MouniraHarzallah, IUT de Nantes - Laboratoire LINA
Nathalie Hernandez, Université de Toulouse Le Mirail - Laboratoire IRIT
Antoine Isaac, Europeana&VrijeUniversiteit Amsterdam
Marie-Christine Jaulent, INSERM - UMR_S 872 EQ20
Clément Jonquet, Université de Montpellier - Laboratoire LIRMM
Gilles Kassel, Université de Picardie Jules Verne - Laboratoire MIS
Khaled Khelif, Cassidian
Pascale Kuntz, Université de Nantes - Laboratoire LINA
Philippe Laublet, Université Paris Sorbonne - Laboratoire STIH
Florence Le Ber, ENGEES - Laboratoire ICube
Michel Leclère, Université Montpellier 2 - Laboratoire LIRMM
Alain Léger, FT R&D - Orange Labs
Jean Lieber, Université Nancy 1 - Laboratoire LORIA
Moussa Lo, Université Gaston Berger, Sénégal
Vanda Luengo, Université Joseph Fourier - Laboratoire LIG
Jean-Charles Marty, Université de Savoie - Laboratoire LIRIS
Nada Matta, Université de Technologie de Troyes - Laboratoire Tech-CICO
Alain Mille, Université Claude Bernard Lyon 1 - Laboratoire LIRIS
Pascal Molli, Université de Nantes - Laboratoire LINA
Alexandre Monnin, INRIA - Wimmics
Amedeo Napoli, CNRS - Laboratoire LORIA
Jérôme Nobécourt, Université Paris 13 - Laboratoire LIM&BIO
Alexandre Passant, seevl
Nathalie Pernelle, Université Paris 11 - Laboratoire LRI
Frédéric Precioso, Université Nice Sophia Antipolis - Laboratoire I3S
Yannick Prié, Université de Nantes - Laboratoire LINA
Sylvie Ranwez, Ecole des Mines d'Ales - Laboratoire LGI2P
Chantal Reynaud, Université Paris-Sud - Laboratoire LRI
Catherine Roussey, Irstea
Pascal Salembier, Université de Technologie de Troyes - Laboratoire Tech-CICO
FrancoisScharffe, Université de Montpellier 2 - Laboratoire LIRMM
Karim Sehaba, Université Lumière Lyon 2 - Laboratoire LIRIS
Milan Stankovic, Sépage& Université Paris-Sorbonne
Sylvie Szulman, Université Paris 13 - Laboratoire LIPN
Eddie Soulier, Université de Technologie de Troyes - Laboratoire Tech-CICO
Andrea Tettamanzi, Université Nice Sophia Antipolis - Laboratoire I3S
Yannick Toussaint, INRIA Nancy Grand-Est
Francky Trichet, Université de Nantes - Laboratoire LINA
Cassia Trojahn, Université de Toulouse Le Mirail - Laboratoire IRIT
Raphaël Troncy, EURECOM
Pierre-Yves Vandenbussche, MONDECA
Serena Villata, INRIA Sophia Antipolis
Amel Yessad, Université Paris 6 - Laboratoire LIP6
Manuel Zacklad, CNAM - Laboratoire DICEN
HaifaZargayouna, Université Paris 13 - Laboratoire LIPN
Antoine Zimmermann, École Nationale Supérieure des Mines de Saint-Étienne
Pierre Zweigenbaum, CNRS - Laboratoire LIMSI

Comité d'organisation

Présidente du comité d'organisation :
Catherine Roussey, Chargée de recherche à l'Irstea

IRIT
Fabien Amarger,
Catherine Comparot
Ollivier Haemmerlé
Nathalie Hernandez
Camille Pradel
Cassia Trojahn Dos Santos

Irstea
Jean Pierre Chanet
Eva Lambert
Irène Mingot
Francois Pinet
Catherine Roussey
Eliane Simon
Vincent Soulignac
Anais Wermeille

LIMOS
Diyé Dia
Nader Jelassi
Yannick Loiseau
Engelbert Mephu Nguifo
Christophe Rey

sémantique
web
ontologie
connaissances
données
construction
utilisateurs
partir
langage
plateforme
sparql
définition
apprentissage
maladies
peuplement
numérique
décision
interface
santé
aide
sein
vers
vue
utilisation
réseau
univers
qualité
exploiter
modulaire
localisées
système
démonstrations
lexicales
terminologies
socio-sémantique
scientifiques
assistance
générique
identifier
partage
fouille
catalogue
création
cuisine
exemple
basée
méthodes
infrastructure
recommandations
automatique
exploitation
communautés
bayésien
cible
feedbacks
situations
naturelle
quels
pilotage
langue
modèle
extension
fma
dont
crise
annotation
prétends
activités
folksonomies
clés
infectieuses
urgences
social
nash
points
indicateurs
veille
programmée
pédagogiques
interopérabilité
rdf
cas
théorique
applications
coopération
patients
cartographie
gérant
intelligente
intégration
thérapeutique
physiopathologie
épiphytes
utilisant
organisationnel
liens
posters
cancer
sens-texte
gérez
documents
prototype
algorithme
patrons
template
genre
évolution
plus
sonore
pretty
collaborative
personnalisées
définitions
être
intérêt
acteurs
noyau
ido-core
recherche
territoire
réutilisation
ontologies
décrire
forums
programmer
iconique
support
organisation
confrontation
usages
diffusez
émotions
théorie
suis-je
information
centré
formelles
point
prise
wiki-based
printing
celui
liées
adaptation
agrégation
gestion
médical
wiki/ide

Avant-propos

Les Journées Francophones d'Ingénierie des Connaissances fêtent cette année leurs 25 ans. Cette conférence est le rendez-vous annuel de la communauté française et francophone qui se retrouve pour échanger et réfléchir sur des problèmes de recherche qui se posent en acquisition, représentation et gestion des connaissances.

Parmi les vingt et un articles sélectionnés pour publication et présentation à la conférence, un thème fondateur de l'ingénierie des connaissances domine : celui de la modélisation de domaines. Six articles traitent de la conception d'ontologies, trois articles de l'annotation sémantique et du peuplement d'ontologies et deux articles de l'exploitation d'ontologies dans des systèmes à base de connaissances. L'informatique médicale est le domaine d'application privilégié des travaux présentés, que l'on retrouve dans sept articles.

L'ingénierie des connaissances accompagne l'essor des technologies du web sémantique, en inventant les modèles, méthodes et outils permettant l'intégration de connaissances et le raisonnement dans des systèmes à base de connaissances sur le web. Ainsi, on retrouve les thèmes de la représentation des connaissances et du raisonnement dans six articles abordant les problématiques du web de données : le liage des données, leur transformation et leur interrogation ; la représentation et la réutilisation de règles sur le web de données ; la programmation d'applications exploitant le web de données.

L'essor des sciences et technologies de l'information et de la communication, et notamment des technologies du web, dans l'ensemble de la société engendre des mutations dans les pratiques individuelles et collectives. L'ingénierie des connaissances accompagne cette évolution en plaçant l'utilisateur au cœur des systèmes informatiques, pour l'assister dans le traitement de la masse de données disponibles. Quatre articles sont dédiés aux problématiques du web social : analyse de réseaux sociaux, détection de communautés, folksonomies, personnalisation de recommandations, représentation et prise en compte de points de vue dans la recherche d'information. Deux articles traitent de l'adaptation des systèmes aux utilisateurs et de l'assistance aux utilisateurs et deux autres de l'aide à la prise de décision.

Le taux de sélection de cette édition de la conférence est de 50%, avec dix-neuf articles longs et deux articles courts acceptés parmi quarante-deux soumissions. S'y ajoutent une sélection de neuf posters et démonstrations parmi douze soumissions, présentés dans une session dédiée et inclus dans les actes. Enfin, une innovation de cette édition 2014 de la conférence est la programmation d'une session spéciale « Projets et Industrie », animée par Frédérique Segond (Viseo), à laquelle participeront Laurent Pierre (EDF), Alain Berger (Ardans) et Mylène Leitzelman (Mnemotix).

Trois conférencières invitées ouvriront chacune des journées de la conférence que je remercie chaleureusement de leur participation. Nathalie Aussenac-Gilles (IRIT) retracera l'évolution de l'ingénierie des connaissances en France depuis 25 ans, de la pénurie à la surabondance. A sa suite, Frédérique Segond (Viseo) abordera le problème de « l'assouvissement » de la faim de connaissances dans la nouvelle ère des connaissances dans laquelle nous sommes entrés. Enfin, Marie-Laure Mugnier (LIRMM) présentera un nouveau cadre pour l'interrogation de données basée sur une ontologie, fondé sur des règles existentielles.

Je remercie vivement les auteurs pour leurs contributions, le comité de programme pour le nombre de ses relectures (quatre par article soumis) et la qualité de celles-ci qui ont contribué à celle des articles présentés, le comité d'organisation de la conférence pour l'efficacité de son travail et tout particulièrement sa présidente, Catherine Roussey avec qui cela a été un plaisir de collaborer, et l'AFIA pour son soutien à la conférence. Je remercie enfin très chaleureusement le bureau du collège IC de l'AFIA et tout particulièrement Nathalie Aussenac-Gilles, Jean Charlet et Fabien Gandon qui m'ont aidée à mener à bien ma mission de présidente du comité de programme d'IC 2014.

Catherine Faron Zucker
Présidente du comité de programme

Sommaire

Construction, peuplement et exploitation d'ontologies.....13

Décrire les maladies localisées et la physiopathologie pour une ontologie des urgences : un algorithme générique à partir de la FMA..... 15
Jean Charlet, Laurent Mazuel, Gunnar Declerck, Patrick Miroux et Pierre Gayet

Construction d'une ontologie modulaire pour l'univers de la cuisine numérique..... 27
Sylvie Despres

IDOSCHISTO : une extension de l'ontologie noyau des maladies infectieuses (IDO-Core) pour la schistosomiase..... 39
Gaoussou Camara, Sylvie Despres et Moussa Lo

Validation de la sémantique d'un langage iconique médical à l'aide d'une ontologie : méthodes et applications..... 51
Jean-Baptiste Lamy, Lina F. Soualmia, Alain Venot et Catherine Duclos

L'intérêt des patrons dans la gestion des connaissances liées à la création sonore……... 63
Antoine Vincent, Bruno Bachimont et Alain Bonardi

Plateforme d'interopérabilité sémantique gérant les terminologies d'interface au sein d'un territoire de santé numérique ……..… 75
Lamine Traore, Amina Chniti, Sajjad Hussain, Nicolas Griffon, Stefan Darmoni, Jean Charlet, Eric Sadou, David Ouagne, Eric Lepage et Christel Daniel

Peuplement automatique d'ontologie à partir d'un catalogue de produits ……. 87
Céline Alec, Brigitte Safar, Chantal Reynaud, Zied Sellami et Uriel Berdugo

Un modèle d'annotation sémantique centré sur les utilisateurs de documents scientifiques : cas d'utilisation dans les études genre ……. 99
Hélène de Ribaupierre et Gilles Falquet

Utilisateurs et usages..... 105

Quels sont les patients atteints d'un cancer du sein dont la décision de prise en charge thérapeutique bénéficie de l'utilisation d'un système d'aide à la décision ? Un exemple utilisant la fouille de données et Onco-Doc2.……... 107
Jacques Bouaud, Arnaud Soulet, Jean-Philippe Spano, Jean-Pierre Lefranc, Isabelle Cojean-Zelek, Brigitte Blaszka-Jaulerry, Laurent Zelek, Axel Durieux, Christophe Tournigand, Nizar Messai, Alexandra Rousseau et Brigitte Seroussi

De la qualité de la coopération à l'identification d'indicateurs de pilotage …… 119
Christopher Couthon, Régis Martineau et Pascal Salembier

Apprentissage de connaissances d'adaptation à partir des feedbacks des utilisateurs …… 125
Abir Beatrice Karami, Karim Sehaba et Benoît Encelle

aLDEAS : un langage de définition de systèmes d'assistance épiphytes ……..… 137
Blandine Ginon, Stéphanie Jean-Daubias, Pierre-Antoine Champin et Marie Lefevre

Web social..... 149

Organisation de communautés et équilibre de Nash ……... 151
Michel Crampes et Michel Plantié

Identifier la cible des émotions dans les forums de santé ……..… 163
Sandra Bringay, EricKergosien, Pierre Pompidor et Pascal Poncelet

Construction et évolution de connaissances par confrontation de points de vue : prototype pour la recherche d'information scientifique ……... 175
Guillaume Surroca, Philippe Lemoisson, Clément Jonquet et Stefano A. Cerri

Vers des recommandations plus personnalisées dans les folksonomies ……. 187
Mohamed Nader Jelassi, Sadok Ben Yahia et Engelbert Mephu Nguifo

Web sémantique..... 199

Swip : une interface langue naturelle à SPARQL programmée en SPARQL..... 201
Camille Pradel, Ollivier Haemmerlé et Nathalie Hernandez

SPARQL Template: un langage de pretty printing pour RDF..... 213
Olivier Corby et Catherine Faron-Zucker

Définition de la sémantique des clés dans le web sémantique : un point de vue théorique..... 225
Michel Chein, Madalina Croitoru, Michel Leclère, Nathalie Pernelle, Fatiha Saïs et Danai Symeonidou

Publication, partage et réutilisation de règles sur le web de données..... 237
Oumy Seye, Catherine Faron Zucker, Olivier Corby et Alban Gaignard

Programmer le web de données avec un « wiki-based IDE »..... 249
Pavel Arapov, Michel Buffa et Amel Ben Othmane

Posters et démonstrations..... 261

Une plateforme support à l'apprentissage organisationnel..... 263
Ala Atrash, Marie-Hélène Abel et Claude Moulin

Infrastructure web socio-sémantique pour la veille collaborative 267
Jean-Pierre Cahier, Mylène Leitzelman et Patrick Brébion

Suis-je celui que je prétends être ? 271
Diyé Dia, Olivier Coupelon, Yannick Loiseau et Olivier Raynaud

Agrégation pour la réparation de liens 275
Léa Guizol

Un éditeur de définitions formelles pour les connaissances lexicales de la théorie Sens-Texte 279
Maxime Lefrançois, Fabien Gandon, Alain Giboin et Romain Gugert

Un wiki/IDE pour exploiter le web de données 283
Pavel Arapov, Amel Ben Othmane et Michel Buffa

Adnosco : gérez les données que vous diffusez ! 287
Nadia Bennani, Emmanuel Gaude, Elöd Egyed-Zsigmond et Philippe Lamarre

Vers une cartographie participative basée sur la communication transversale des acteurs dans les situations de crise 291
Amina Saoutal, Jean-Pierre Cahier et Nada Matta

Intégration d'un réseau bayésien dans une ontologie……. 295
Emna Hlel, Salma Jamoussi et Abdelmajid Ben Hamadou

Index des auteurs.....299

Construction, peuplement et exploitation d'ontologies

Décrire les maladies localisées et la physiopathologie pour une ontologie des urgences : un algorithme générique à partir de la FMA

Jean Charlet[1,2], Laurent Mazuel[2,3], Gunnar Declerck[2,4], Patrick Miroux[5], Pierre Gayet[6]

[1] Assistance Publique – Hôpitaux de Paris, F-75004, Paris, France ;

[2] INSERM, U1142, LIMICS, F-75006, Paris, France, Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1142, LIMICS, F-75006, Paris, France, Université Paris 13, Sorbonne Paris Cité, LIMICS, (UMR_S 1142), F-93430, Villetaneuse, France;

[3] ANTIDOT, F-13410, Lambesc, France ;

[4] Costech, Université de Technologie de Compiègne, F-60200, Compiègne, France ;

[5] Dpt Urgences CHU d'Angers, F-49100, Angers, France ;

[6] Centre hospitalier de Compiègne, F-60200, Compiègne, France.

Résumé :
ONTOLURGENCES est une représentation termino-ontologique du domaine de la médecine d'urgence initialement développée pour la recherche d'information. Cette ressource est conçue pour assurer *a*) le rôle de modèle du domaine répertoriant tous les concepts pertinents et *b*) le lien entre les concepts et la façon dont ils sont nommés dans les documents du Dossier Patient Informatisé. Cette double fonction permet l'annotation et l'indexation de dossiers patients et assure une recherche d'informations rapide. Le développement d'ONTOLURGENCES nous a amené à nous intéresser à l'articulation des points de vue physiologiques et anatomiques, deux composantes essentielles des modèles en médecine. Pour gérer les difficultés liées à une modélisation croisée de ces deux points de vue, très rapidement complexe et source d'erreur, il nous est apparu nécessaire d'exploiter une ressource de référence, le *Foundational Model of Anatomy*, *la FMA*, afin de réaliser une modélisation systématique de l'anatomie. Nous avons ainsi développé une méthode qui permet de générer de manière semi-automatique, à partir de la FMA : la branche des maladies localisés – *e.g. maladie du tube digestif* –, la branche de l'anatomie – *e.g.* le *tube digestif* lui-même – et, enfin, la définition des axiomes de localisation – *e.g.* le fait qu'une *maladie du tube digestif* est localisée sur le *tube digestif*. Les premiers résultats permettent de valider la pertinence de notre démarche.

Mots-clés : Ontologie médicale, épistémologie de la modélisation, anatomie médicale, physiopathologie médicale.

1 Introduction

Le projet LERUDI (LEcture Rapide en Urgence du Dossier patient Informatisé) vise à développer un système d'information (SI) offrant aux professionnels de santé une vision synthétique du dossier patient informatisé, et la possibilité d'un parcours rapide de celui-ci, pour permettre des prises de décisions médicales soumises à d'importantes contraintes de temps. Le champ d'expérimentation de ce projet est la lecture d'un dossier hospitalier par un médecin urgentiste. LERUDI est fondé sur une Ressource Termino-ontologique (RTO) [1] nommée ONTOLURGENCES, qui assure *a*) le rôle de modèle du domaine répertoriant tous les concepts pertinents

1. Une RTO se définit comme une ontologie dans laquelle les termes sont rattachés aux concepts de façon systématique et exhaustive. Il y a plusieurs façons de rattacher des termes à des concepts selon ce qu'on veut pouvoir représenter (Reymonet, 2007; Vandenbussche & Charlet, 2009).

et *b*) le lien entre les concepts et la façon dont ils sont nommés dans les documents du dossier patient. Cette double fonction doit permettre l'annotation et l'indexation de dossiers patients et la recherche d'information (RI) dans les dossiers indexés.

Les différentes étapes du développement de cette RTO ont été décrites dans (Charlet *et al.*, 2012b). Dans le présent article, nous allons nous intéresser à l'articulation, au sein de cette RTO, des points de vue physiopathologiques et anatomiques. Ces deux points de vue jouent un rôle essentiel dans la caractérisation des pathologies en médecine : par définition, une pathologie est un processus biologique anatomiquement localisé, par sa cause ou ses manifestations observables. Le travail de construction d'une ontologie médicale doit donc, d'une façon ou d'une autre, les intégrer et les modéliser explicitement. Or, cette entreprise pose d'importantes difficultés. Si, de prime abord, les grands concepts de l'anatomie (le corps humain et son découpage en éléments plus fins) et ceux de la physiopathologie [2] sont accessibles, leur modélisation précise devient très rapidement complexe et source d'erreurs. Et la nécessité d'assurer la cohérence (logique aussi bien que sémantique) des modélisations de chacun des deux points de vue rend la tâche plus complexe encore.

Pour remédier à ces difficultés, nous avons fait le choix d'utiliser une ressource de référence : le *Foundational Model of Anatomy* (*la FMA* dans la suite de cet article) pour réaliser le travail de modélisation systématique de l'anatomie dans ONTOLURGENCES. Nous avons ainsi développé une méthode qui permet de réutiliser l'énorme travail de modélisation de la FMA pour générer de manière semi-automatique la branche des maladies localisés – *e.g. maladie du tube digestif* –, la branche de l'anatomie – *e.g.* le *tube digestif* lui-même – et, enfin, la définition des axiomes de localisation – *e.g.* le fait qu'une *maladie du tube digestif* est localisée sur le *tube digestif*.

Dans la section 2, nous présentons la problématique de modélisation à laquelle nous sommes confrontés ; dans la section 3, nous décrivons les deux ressources impliquées dans notre projet, l'ontologie ONTOLURGENCES et la FMA ; dans la section 4 nous présentons la méthodologie proposée et l'algorithme développé ; dans la section 5, nous donnons les premiers résultats de ce travail ; enfin, dans la section 6, nous concluons et proposons quelques perspectives.

2 Modélisation de la physiopathologie *vs* de l'anatomie

La méthode anatomoclinique a été fondée par l'école française (Leannec, Bretoneau, Trousseau) dans la première moitié du XIXe siècle. Elle consiste en la mise en relation de la clinique et de l'anatomie. Cette base de compréhension des maladies servira à leur traitement. Pendant la seconde moitié du XIXe siècle, la découverte de la nature microbiologique des infections (Pasteur) et l'invention de la médecine expérimentale (C. Bernard) impulsent une analyse des mécanismes des maladies : la physiopathologie.

Depuis, l'anatomie, la physiologie et la physiopathologie sont enseignées en médecine et forment le cœur des compétences des médecins. Si l'on cherche à modéliser les connaissances de base d'une spécialité – ici les Urgences – on est rapidement confronté à la nécessité de modéliser l'anatomie et la physiopathologie de la médecine dans les différents champs d'intervention. Par exemple, dans les urgences, on va s'intéresser aux fractures, du point de vue de l'anatomie – l'os – et de la physiopathologie – un traumatisme ou une métastase.

2. On organise habituellement les processus physiopathologiques selon 6 classes : les processus obstructifs, traumatiques, inflammatoire et infectieux, dégénératifs, tumoraux et enfin psychopathologiques.

C'est pourquoi les ontologies en médecine clinique modélisent l'anatomie et la physiopathologie. Comme indiqué en introduction, cette modélisation n'est pas aisée, en particulier en raison : 1) de la taille du domaine en termes de concepts et 2) de la nécessaire interaction de ces 2 points de vues.

Ainsi, un cancer doit être caractérisé par son type de cellules mais doit également pouvoir être localisé. Pour faciliter la conception d'ONTOLURGENCES et en améliorer l'interopérabilité, nous avons cherché à réutiliser différentes ressources du domaine biomédical (Charlet *et al.*, 2012b), notamment la branche des diagnostics de la SNOMED V3.5 La modélisation proposée par la SNOMED V3.5 s'est toutefois rapidement montrée insuffisante, car trop imprécise (granularité insuffisante), mais surtout mal organisée sur le plan de la physiopathologie – ce qui nous a passablement surpris. Des 25000 maladies répertoriées dans la SNOMED V3.5, seules 6500 ont ainsi été conservées (Charlet *et al.*, 2012b).

La cohérence générale d'ONTOLURGENCES à l'issu de ce travail de modélisation nous a toutefois laissé insatisfaits. La physiopathologie était plus ou moins bien modélisée, mais surtout l'anatomie était extrêmement lacunaire, en raison du nombre de concepts impliqués : par exemple, si l'on s'intéresse aux fracture des os, une modélisation satisfaisante sur un plan sémantique et opératoire (utilisable pour la RI avec de bonnes performances) exige : (a) de modéliser l'ensemble des os et leur topographie interne (tête, etc.), et (b) de définir chaque concept de fracture par une relation sémantique référant à l'os associé (ou aux os s'il s'agit d'une fracture multiple ou d'une structure osseuse complexe). Ce travail, extrêmement lourd, doit être répété pour toutes les structures anatomiques impliquées dans les pathologies auxquelles sont susceptible d'avoir affaire les urgentistes. Dans ce contexte, la FMA était bien, par son degré de détail et d'exhaustivité, la référence attendue. Mais inversement, précisément en raison de sa taille, le nombre d'éléments à réutiliser et les difficultés pour s'y orienter rendaient sa consultation et sa réutilisation durant la modélisation problématique, et partiellement aléatoire.

A ces difficultés, s'ajoutent des problèmes de choix de modélisation inhérents à toute conception d'ontologie. Notamment, il y a aujourd'hui un certain consensus sur l'idée qu'une modélisation ontologique bien faite (i) privilégie un axe de différentiation qui sert à organiser les relations de subsomption et (ii) modélise les autres axes via les relations (Bouaud *et al.*, 1995; Bachimont *et al.*, 2002). Ce qui signifie qu'on n'autorise qu'une seule relation de subsomption assertée par concept (pas de polyparentalité). Le premier axe sera constitutif de la hiérarchie de l'ontologie, définissant l'arbre *is-a*. Le second interviendra lors de la classification des concepts définis (Horridge, 2009). Dans ce contexte, nous avions donc l'alternative suivante :

> (1) utiliser le point de vue physiopathologique pour définir l'arbre is-a de notre ontologie et reconstruire le point de vue anatomique par classification, ou :
>
> (2) utiliser le point de vue anatomique pour définir l'arbre is-a et reconstruire le point de vue physiopathologique par classification.

Différentes raisons nous ont amené à adopter la première approche. D'abord par rapport à la physiopathologie : 1) la physiopathologie n'est généralement pas très bien maîtrisée du point de vue de la modélisation ontologique ; 2) corollaire, il n'y a pas de ressource de référence à ce sujet (notre expérience avec la SNOMED V3.5 nous l'a bien montré) [3] ; 3) la relation phy-

3. Notre raisonnement à ce sujet est contre-intuitif : puisque la modélisation physiopathologique est difficile à faire et qu'il n'y a donc pas d'ontologie satisfaisante en ce domaine, prenons les moyens de construire cette partie du modèle et, à l'inverse, pour l'anatomie, réutilisons une ressource (la FMA).

siopathologique constitue souvent l'articulation entre les différentes pathologies des patients complexes. Or les pathologies chroniques constituent un enjeu de santé publique actuel, y compris aux urgences [4], 4) enfin, en termes de concepts à modéliser, la physiopathologie représente une dimension bien plus petite que l'anatomie. 5) l'anatomie est assez bien maîtrisée du point de vue de la modélisation ontologique et cela explique, entre autres, l'existence de la FMA ; 6) parce que la FMA est disponible, on peut imaginer l'utiliser de façon plus ou moins automatique pour constituer la branche anatomique d'ONTOLURGENCES, 7) enfin, pour revenir à la question de l'essence unique de la modélisation, pour beaucoup de pathologies, il existe un positionnement physiopathologique unique alors que celles-ci ont un positionnement anatomique multiple (bronchopneumopathie).

L'essentiel de notre travail de modélisation a donc consisté à analyser aussi précisément que possible le point de vue physiopathologique tel qu'il est représenté dans la médecine d'urgences et à exploiter la FMA pour reconstruire par classification automatique la branche de l'anatomie dans son entier. En termes plus précis, cela revient à :

1. organiser l'ensemble des concepts de maladies modélisés par l'ontologie dans un arbre de subsomption représentant la perpective physiopathologique telle que la comprend la médecine d'urgence – *e.g. AnevrismeCardiaque is-a Anevrisme* – ;
2. développer la branche des maladies localisés avec des concepts définis (pour que puissent être positionnés sous ces derniers les concepts définis en (1) par classification automatique) – *e.g. MaladieCardiaque is-a MaladieDUnOrgane* –, ce qui implique de :
3. définir les axiomes de localisation – *e.g.* le fait qu'une maladie du cœur est localisée sur le cœur : MaladieDUnOrgane and (localized some Heart) and (localized only Heart)–, et enfin,
4. développer la branche de l'anatomie nécessaire à l'expression des axiomes – *e.g.* le *cœur* lui-même : Heart is-a OrganWithCavitatedOrganParts is-a Organ.

3 Ressources utilisées

Les ressources utilisées sont donc l'ontologie des urgences – ONTOLURGENCES– dans une version en perpétuelle évolution à partir de la 3.0.3 et la version OWL de la FMA.

ONTOLURGENCES v3.03. ONTOLURGENCES sert de test à ce travail depuis sa version de juillet 2013. Elle avait à ce moment 10191 classes, 60 *Object Properties*, 1 *Data property* et 11591 axiomes logiques dont 11339 axiomes de sous-classe et 89 axiomes de classes d'équivalence. L'ontologie est construite sous l'ontologie noyau de OntoMénélas [5] (Charlet *et al.*, 2012a).

FMA. La FMA (*Foundational Model of Anatomy*) est « une ontologie de référence sur l'anatomie humaine » (Rosse & Mejino, 2003; Rosse & Jr, 2008). Elle est destinée, d'après ses concepteurs, à modéliser l'anatomie humaine canonique, c'est-à-dire « l'anatomie idéale à laquelle chaque individu et ses parties doivent se conformer » (Rosse & Jr, 2008). Elle contient plus de 85 000 classes, 140 relations reliant les classes, et plus de 120 000

4. http://www.ameli.fr/fileadmin/user_upload/documents/cnamts_rapport_charges_produits_2014.pdf

5. http://purl.oclc.org/NET/spim/ontologies/public/OntoMenelas/

termes. La plupart des entités sont des structures anatomiques composées de plusieurs parties interconnectées de façon complexe, décrites, par exemple, en termes de topographie, constituants, innervation, vaisseaux sanguins, limites ou frontières. Par exemple, un cœur a deux régions (le côté gauche et le côté droit), des parties le constituant (par exemple, la paroi du cœur, les diverses valves, le réseau d'innervation cardiaque, le septum interventriculaire, le septum auriculo-ventriculaire, les cavités, etc.), Il est innervé par le plexus cardiaque profond et alimenté par les artères coronaires, etc. (Golbreich *et al.*, 2013). De l'avis de tous, et malgré quelques bogues et insuffisances dans le modèle, elle est reconnue comme l'ontologie de référence de l'anatomie humaine.

Puisque la FMA a d'abord été développée dans un langage de *frames*, sa disposition en tant que fichier OWL, et *a fortiori* OWL2, ressort de travaux spécifiques de l'équipe de développment de la FMA comme d'autres chercheurs (Golbreich *et al.*, 2013). La version que nous avons utilisée est la version 3.2.1 en OWL Full de la FMA disponible sur le site en téléchargement [6] : elle a 84 454 classes, 237 382 instances, 132 *Object Properties*, 167 *Data property* et 1 719 576 axiomes dont 87803 axiomes de sous-classe. Le grand nombre d'axiomes et d'instances sont liés au caractère *full* de cette version de la FMA où chaque classe est à la fois instance et sous-classe de la classe mère.

4 Méthode de génération

4.1 Notations et vocabulaire

Dans la suite de cet article, nous utiliserons les notations suivantes :

— $\mathcal{O}$ l'ontologie de travail initiale et $\mathcal{O}_{fma}$ la modélisation OWL de la FMA.

— Nous noterons l'espace de nom de l'ontologie $\mathcal{O}$ « `onto:` » et celui de la FMA « `fma:` ». Ainsi, « `onto:Cancer` » et « `fma:Heart` » sont respectivement des URIs valides dans l'ontologie $\mathcal{O}$ et dans la FMA $\mathcal{O}_{fma}$.

— Nous considérerons qu'une ontologie est définie par l'ensemble des axiomes la constituant. nous utiliserons ainsi la notation ensembliste appliquée à des axiomes (énoncés en écriture fonctionnelle de OWL2 [7]) pour parler de l'appartenance d'objet à l'ontologie. Par exemple : $Declaration(Class(\texttt{fma:Heart})) \in \mathcal{O}_{fma}$.

Nous définissons aussi quelques fonctions utilitaires qui seront utilisés dans les versions plus formelles des algorithmes de l'article [8] :

— « `fma_concept` » est une fonction qui permet de récupérer la Classe FMA associée à un identifiant numérique FMA. Pour cela, nous utilisons l'annotation « `fma:FMAID` » présente dans la FMA :

$$fma\_concept(fma\_id) = \{cpt | AnnotationAssertion(\texttt{fma:FMAID}\ cpt\ fma\_id) \in \mathcal{O}_{fma}\}_1 \quad (1)$$

6. `http://sig.biostr.washington.edu/projects/fma/release/v3.2.1/alt_formats.html`

7. http://www.w3.org/TR/owl2-syntax/

8. Certaines de ces fonctions sont suffixées par un « 1 » car nous prenons le premier élément pour éviter que le type de retour soit un ensemble. En effet, même si la théorie autorise plusieurs annotations, en pratique la FMA ne définit qu'un seul label et un seul identifiant numérique pour chaque concept.

— « `fma_label` » est une fonction qui permet de récupérer le label d'un concept FMA étant donné sa Classe. Pour cela, nous utilisons simplement le `rdfs:label` :

$$fma\_label(fma\_concept) = \{str | AnnotationAssertion(\texttt{rdfs:label}\ fma\_concept\ str) \in \mathcal{O}_{fma}\}_1 \quad (2)$$

— « `fragment` » renvoie le fragment de l'URI, c'est-à-dire l'URI sans espace de nom. Par exemple $fragment(\texttt{fma:Heart}) ='' Heart''$

4.2 Hypothèses initiales sur l'ontologie

Puisque la motivation principale est l'enrichissement automatique d'une ontologie médicale ayant besoin de localisation anatomique au moyen de la modélisation de la FMA, l'étape initiale est l'annotation de chaque concept de l'ontologie par un identifiant FMA.

Dans notre ontologie $\mathcal{O}$, nous avons donc ajouté l'annotation `onto:pourFMA` qui pour chaque concept de $\mathcal{O}$ permet d'associer l'identifiant numérique unique d'un concept de la FMA. Cette étape est faite manuellement par les experts médecins du domaine de l'ontologie. A la demande des médecins, nous avons autorisé la multi-annotation (*i.e.* permettre d'associer plusieurs localisations à un seul concept). Cette multi-localisation se traduira alors par une intersection. Nous ne traitons pas actuellement de situation d'union entre plusieurs localisations.

L'algorithme que nous présentons à la section suivante suppose ainsi que tous les concepts ayant besoin d'une localisation sont annotés par une ou plusieurs annotations `onto:pourFMA`.

4.3 Algorithme de construction

4.3.1 Présentation générale

L'algorithme que nous présentons possède 3 objectifs majeurs :
— recopier la partie de la FMA pertinente dans l'ontologie ;
— construire une arborescence des maladies localisées ;
— ajouter des restrictions some/only entre les concepts de l'ontologie initiale.

Ceci nous permet à terme d'effectuer des inférences sur l'ontologie pour classifier automatiquement les concepts originaux sous les nouvelles maladies (e.g. inférer que « Cancer du poumon » est une « maladie du poumon » grâce à la relation de localisation sur « poumon » de la FMA). Les trois sections suivantes présentent les trois parties de l'algorithme.

4.3.2 Recopie de la FMA

C'est la partie la plus simple de l'algorithme. Pour chaque concept possédant une annotation `onto:pourFMA`, nous recopions dans $\mathcal{O}$ l'ensemble des axiomes d'annotations du concept de la FMA ainsi que tous ses concepts parents jusqu'au concept `fma:Physical_anatomical_entity`[9]. Une fois ce concept atteint, la hiérarchie doit être stockée dans l'ontologie d'origine puisque les concepts anatomiques doivent être accessibles pour construire les restrictions de localisation (voir § 4.3.3). Pour ONTOLURGENCES qui représente notre cas d'étude, nous l'avons

9. La hiérarchie de la FMA hors de cette partie est abstraite et ne présente pas d'intérêt de localisation.

stocké sous `onto:StructureAnatomique`, mais le concept exact est bien entendu dépendant de l'ontologie dans laquelle la FMA est fusionnée.

A noter que nous conservons l'URI original de la FMA afin de garder tous les avantages du Linked Data. La recopie en interne ne sert qu'à des fins de visualisation et d'inférence hors ligne sur des logiciels tels que Protégé. Nous ne faisons que des recopies et ne procédons à aucune modification de la sémantique initiale de la FMA.

4.3.3 Création de la hiérarchie des maladies

L'idée générale est de construire un concept « DiseaseOf » pour chaque localisation recopiée de la FMA. Par exemple, si l'on recopie `fma:Heart`, alors nous allons créer un concept `onto:DiseaseOfHeart`. Néanmoins, la hiérarchie anatomique de la FMA ne se transpose pas directement en une hiérarchie des maladies. Il faut utiliser la hiérarchie tout-partie pour être pertinent médicalement parlant. Si par exemple un patient a une douleur thoracique, il a une maladie d'un organe du thorax (cœur/poumon). En suivant la subsomption anatomique, on pourrait inférer que la maladie du poumon est une maladie d'un organe. Si, dans l'absolu, cette subsomption n'est pas fausse, elle n'a aucune pertinence médicale donc aucun intérêt. A l'inverse, par une relation tout-partie, on infèrera que la maladie du poumon est une maladie du système respiratoire inférieur ou une maladie du thorax et on sera médicalement pertinent.

Pour cela, nous allons utiliser les relations tout-partie de la FMA. Elles sont au nombre de 4 (Mejino *et al.*, 2003) :

— `fma:part` / `fma:part_of`
— `fma:regional_part` / `fma:regional_part_of`
— `fma:systemic_part` / `fma:systemic_part_of`
— `fma:constitutional_part` / `fma:constitutional_part_of`

Ces relations sont symétriques, de sorte que $ObjectPropertyAssertion(X\_part\ a\ b) \iff ObjectPropertyAssertion(X\_part\_of\ b\ a)$. En pratique, la représentation OWL de la FMA liste explicitement tous les axiomes et leur contraire [10].

L'algorithme que nous allons utiliser est alors simple :

— à partir d'un concept donné, si il existe une relation « part_of » partant de ce concept, alors nous créons comme père « DiseaseOf » le concept au bout de ce part_of. Par exemple, si nous avons l'axiome $ObjectPropertyAssertion($`fma:part_of fma:Lung fma:Thorax`$)$, alors nous créons l'axiome $SubClassOf($`onto:DiseaseOfLung onto:DiseaseOfThorax`$)$;
— afin de garantir la connexité finale du graphe, si le concept n'a pas de relation « part_of » nous suivons alors la relation de subsomption classique.

Cette nouvelle hiérarchie est branchée sous le concept `onto:Diagnostic`. L'algorithme est défini plus formellement plus bas (voir algorithme 1).

D'autre part, les concepts de la hiérarchie construite à l'étape précédente peuvent être *définis* tel que les restrictions sont des relations vers la FMA. Par exemple « `onto:DiseaseOfLung` » est sémantiquement défini par « une maladie localisée sur une et uniquement une instance de poumon ». Pour cela, nous utilisons la relation « `onto:localisedOn` » définie entre l'arborescence des diagnostics et l'arborescence des localisations. Soit un concept « `fma:X` » et soit « `fma:Y` » le père de « `fma:X` », alors le concept « `onto:DiseaseOfX` » est un concept défini

10. La parité n'est pas strictement exacte dans le fichier, mais nous n'avons pas trouvé de contre-exemple nous laissant supposer que ce soit pour une autre raison que l'absence de complétude de la FMA.

par l'axiome suivant :

$$EquivalentClass(\texttt{onto:DiseaseOfX} \; ObjectIntersectionOf(\texttt{onto:DiseaseOfY} \; ObjectSomeValuesFrom(\texttt{onto:localizedOn fma:X}) \; ObjectAllValuesFrom(\texttt{onto:localizedOn fma:X})))$$

Si l'on instancie l'exemple précédent avec le couple Poumon/Thorax, cela donne :

$$EquivalentClass(\texttt{onto:DiseaseOfLung} \; ObjectIntersectionOf(\texttt{onto:DiseaseOfThorax} \; ObjectSomeValuesFrom(\texttt{onto:localizedOn fma:Lung}) \; ObjectAllValuesFrom(\texttt{onto:localizedOn fma:Lung})))$$

Algorithme 1 Construction d'un concept DiseaseOf

```
fonction build_parent_part(cpt):
  // Avec X ∈ {regional, systemic, constitutional, }
  retourne {r|ObjectPropertyAssertion(X_part r cpt)} U
                                    {r|ObjectPropertyAssertion(X_part_of cpt r)}

fonction build_disease_of(concept_fma):
  label_fma = fma_label(concept_fma)
  new_uri = "onto : DiseaseOf" + fragment(concept_fma)
  insert(O, Declaration(Class(new_uri)))
  insert(O, AnnotationAssertion(rdfs : label new_uri "Disease of " + label_fma)))
  parent_part = build_parent_part(concept_fma)
  si parent_part ≠ ∅
    pour chaque c ∈ parent_part
      si c = fma : Physical_anatomical_entity
        insert(O, SubClassOf(Class(new_uri), onto : Diagnostic)
      sinon
        parent_new_concept = build_disease_of(c)
        insert(O, SubClassOf(Class(new_uri), parent_new_concept)
    retourner Class(new_uri)
  pour chaque c ∈ {p|SubClassOf(concept_fma, c) ∈ O}
    si c = fma : Physical_anatomical_entity
      insert(O, SubClassOf(Class(new_uri), onto : Diagnostic)
    sinon
      parent_new_concept = build_disease_of(c)
      insert(O, SubClassOf(Class(new_uri), parent_new_concept)
```

4.3.4 Ajout de restrictions dans l'ontologie initiale

Afin que les inférences puissent localiser correctement les concepts de l'ontologie initiale, il est nécessaire d'ajouter des restrictions aux concepts de la hiérarchie construite en 4.3.3.

Ces restrictions seront équivalentes aux restrictions utilisées pour les concepts « DiseaseOf », mais les concepts ne seront pas définis (*cf.* infra). D'autre part, nous devons tenir compte du fait que les annotations `onto:pourFMA` peuvent être multiples pour un concept donné (nous considérons cette multiplicité comme une intersection).

Ainsi, pour chaque concept `onto:X` de l'ontologie initiale contenant une annotation `onto:pourFMA` vers les identifiants « id_1 », « id_2 », ..., « id_n », les axiomes suivants seront ajoutés :

$SubClassOf(onto : XObjectSomeValuesFrom(\texttt{onto:localizedOn}\ fma\_concept(id_1))$
...
$SubClassOf(onto : XObjectSomeValuesFrom(\texttt{onto:localizedOn}\ fma\_concept(id_n))$

Et si $n = 1$:

$SubClassOf(onto : XObjectAllValuesFrom(\texttt{onto:localizedOn}\ fma\_concept(id_1))$

Sinon :

$$SubClassOf(\texttt{onto:X}\ ObjectAllValuesFrom(\texttt{onto:localizedOn}\ ObjectIntersectionOf(fma\_concept(id_1)\ \ldots\ fma\_concept(id_n)))$$

4.3.5 Conclusion

L'algorithme final consiste en la juxtaposition des 3 parties précédentes comme indiqué dans l'algorithme 2.

Algorithme 2 Enrichissement global de l'ontologie cible

```
Pour chaque (c, id) ∈ {(c, id)|AnnotationAssertion(onto:pourFMA c id) ∈ O}
  c_fma = fma_concept(id)
  copy(c_fma, O)
  build_disease_of(c_fma)
  add_restrictions(c, c_fma)
```

5 Résultats et discussion

L'ontologie fabriquée a 12396 classes, 60 *Object Properties*, 1 *Data property* et 17072 axiomes dont 13332 axiomes de sous-classe et 3559 axiomes de classes d'équivalence. Ces chiffres sont intéressants à comparer à ceux de l'ontologie dans sa version précédente, en particulier au niveau des axiomes de classes d'équivalence, 89 *vs* 3559 qui traduisent la création de nouveaux concepts définis (voir figure 1).

Un certain nombre de remarques méthodologiques peuvent être faites :

Restriction sur les localisations. Quand on localise une maladie avec l'annotation pourFMA, on a un effet de bord liée à la représentation formelle des classes dans une ontologie. Ainsi, si on annote une maladie du poumon avec l'identifiant du poumon de la FMA puis une maladie plus précise – *e.g.* emphysème pulmonaire – avec le même identifiant,

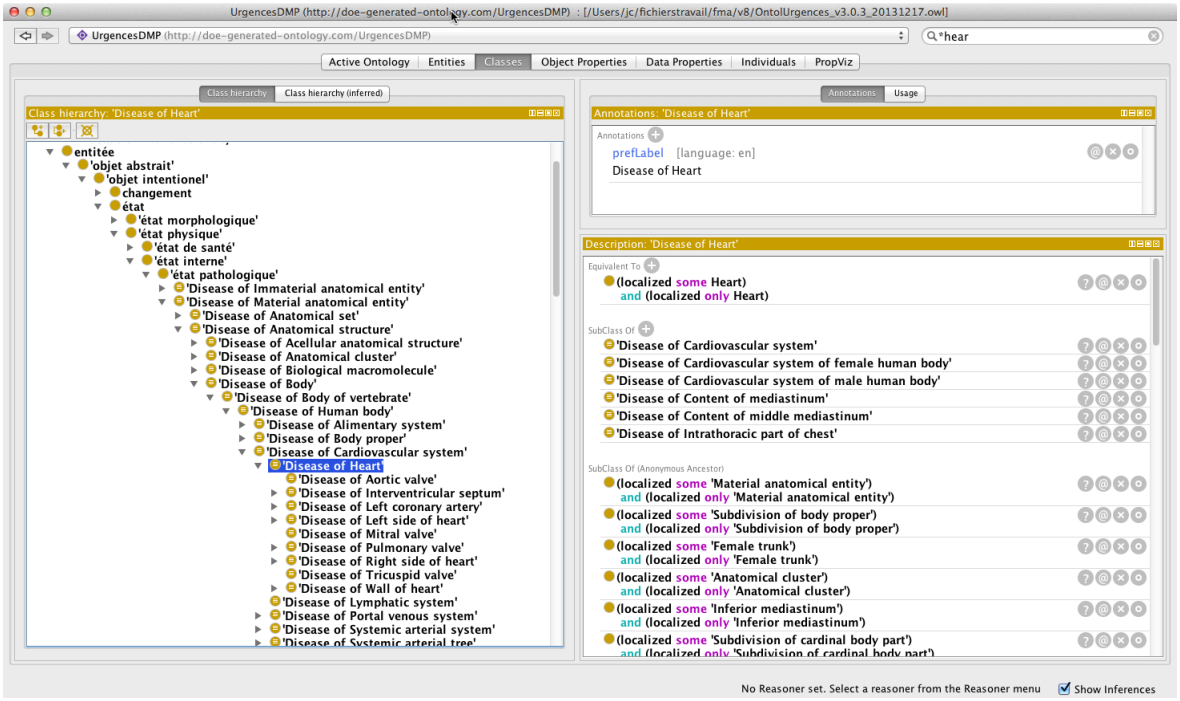


FIGURE 1 – Copie d'écran de l'ontologie vue dans Protégé au niveau du concept « Disease of Heart ». On voit au-dessus l'arborescence des maladies remontant aux concepts de l'ontologie noyau de Onto-Ménélas. Au niveau de la fenêtre des descriptions, on voit la description du concept défini et une grande partie des localisations héritées des concepts parents.

une application simpliste de l'algorithme qui génère les localisations, générerait deux concepts définis équivalents. Pour éviter cela mais pour que cette localisation soit directement attachée à un concept, il faut donc localiser le concept mais seulement en tant que restriction sur un concept primitif. Ainsi ce concept est localisé sur le poumon mais ce n'est pas définitoire, d'autres précisions (restrictions) sont possibles.

Sur la nature et la qualité de la FMA. On a précisé au 5, la version de la FMA que nous utilisons. En fait, ce n'est pas un problème car même s'il y a de nombreuses discussions sur la façon de générer la FMA dans un des différents dialectes (Golbreich *et al.*, 2013) d'OWL – OWL 1 Light, OWL 1 Full, OWL 2 DL, ...–, les liens de tout-partie ne sont pas impactés. Notre seul intérêt en l'espèce est que le dialecte OWL dans lequel la FMA est livrée soit stable pour éviter d'avoir à remettre en place plusieurs API de chargement. Par ailleurs, la FMA n'est pas exempte de bogues – plutôt rares – dont on ne discutera pas ici mais elle a surtout été construite avec des choix de modélisation discutables (au sens propre). Notre but n'est pas de revenir sur les choix de modélisation de la FMA, voire de les corriger car, à ce moment, notre processus de construction de l'ontologie ne pourrait pas être automatisée sur la partie anatomie. Comme, en pratique, la structure des part_of de la FMA est très complexe, nous cherchons plutôt à l'analyser et l'utiliser le plus finement possible.

Utiliser les relations « is-a ». Quand on utilise les part_of pour construire nos maladies localisées, les concepts de plus haut niveau générés ne remontent pas à un concept unique

en raison de la structure de la FMA : les relations « is-a » doivent former un treillis, les autres relations n'apportent pas cette contrainte donc notre algorithme non plus. Or pour des raison de clarté, il serait logique que cela soit le cas. La solution est d'utiliser les relations « is-a » quand on n'a plus rien d'autre. La pertinence médicale n'est alors pas évidente à cet endroit du treillis mais on est de toute façon à des niveaux de généralités pas très médicaux. Par analogie avec les ontologies de haut niveau, on pourrait considérer le haut des maladies localisées, comme un point de vue un peu « philosophique » sur celles-ci. Mais la justification de ce haut est encore à construire. Par ailleurs, on note que le treillis des « Diseaseof » est très dense en arcs par endroit. En analysant rapidement le treillis, on n'a rien vu qui n'ait pas de pertinence médicale mais on peut imaginer de nouvelles analyses pour diminer le nombre d'arcs.

Un algorithme additif. L'un des avantages de notre algorithme est qu'il est incrémental : il est possible de l'appliquer sur une ontologie résultat elle-même de l'algorithme. Autrement dit, il n'est pas nécessaire de garder une version « hors-algorithme » pour pouvoir l'appliquer plusieurs fois, ni de nettoyer manuellement les nouvelles données pour s'en servir. Chaque concept est re-parcouru et les axiomes et restrictions sont créés si necessaires. De la même manière, les arborescences « Disease » et FMA sont réduites afin de ne garder que les concepts utiles (lors de la suppression d'une annotation, il est en effet possible qu'un concept de la FMA deviennent inutilisé et par conséquent que le concept « Disease » associé devienne lui-même inutilisé). Il existe néanmoins une contrainte : si l'utilisateur change (ou supprime) manuellement le litéral d'une annotation « pourFMA » sur un concept donné, alors il doit aussi retirer manuellement les axiomes et restrictions que l'algorithme a créés. En effet, les axiomes ajoutés n'étant pas étiquetés, il est difficile (voir impossible) de faire la différence algorithmiquement entre un axiome ou une restriction ajoutée manuellement est un axiome ou une restriction calculée automatiquement. Pour finir sur ce sujet, on notera que le fait d'avoir des arborescences réduites à ce qui est nécessaire est en accord avec l'idée – défendue par les auteurs – que la FMA est une ontologie de référence dont on prend juste ce qui nous est nécessaire.

La question de la langue des termes. ONTOLURGENCES est une RTO qui sert entre autres à faire de la RI. La version française des termes doit donc être fournie par l'ontologie. Le travail fait ici est en anglais. Pour le faire en français, il faudrait que chacun des éléments anatomiques de la FMA ait un terme français associé. Ce n'est pour l'instant pas le cas. On envisage pour cela de reprendre des traductions faites par le CISMeF sur le portail EHTOP [11] et de les proposer aux auteurs de la FMA pour l'enrichir en termes français.

6 Conclusion

Une première étape de l'évaluation de ce travail a été faite en interne et a d'abord permis de mettre au point l'algorithme exposé ici. La complexité du modèle de la FMA et la compréhension de ce que ça pouvait donner a amené de très nombreux cycles de mise au point. Le premier résultat est que nous avons réellement à notre disposition un algorithme qui nous affranchit d'une tâche longue, fastidieuse et source d'erreurs. Mais cet algorithme nécessite en-

11. `http://www.ehtop.eu/`

core des mises au point et un travail manuel à la fin. La question est aussi de savoir si l'on pourra s'affranchir de ce travail manuel ou le contenir à quelques actions précises terminales.

L'étape suivante de l'évaluation pourrait être de reprendre la nouvelle version de ONTOLURGENCES dans le projet LERUDI et faire une comparaison de l'indexation de quelques dossiers patients informatisés avec l'ancienne version *versus* la nouvelle. Une perspective a plus long terme est la mise en œuvre de cet algorithme sur une ontologie d'un autre domaine médical, qu'elle ait déjà été développée – et là, on suit le même processus que pour ONTOLURGENCES avec des tâches liées au « re-engineering » d'une RTO existante mais que, pour raisons de place, nous n'avons pas développé ici – ou qu'elle soit en cours de développement et la partie anatomique est alors développée par la méthodologie exposée ici. Pour terminer, le travail présenté ici n'est valable que pour la médecine et la FMA mais c'est un choix assumé, en raison du caractère unique de cette ontologie de référence et de l'importance du résultat poursuivi.

Références

BACHIMONT B., ISAAC A. & TRONCY R. (2002). Semantic Commitment for Designing Ontologies : A Proposal. In A. GOMEZ-PÉREZ & V. BENJAMINS, Eds., *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'02)*, volume 2473 of *Lecture Notes in Artificial Intelligence*, p. 114–121, Sigüenza, Espagne : Springer Verlag.

BOUAUD J., BACHIMONT B., CHARLET J. & ZWEIGENBAUM P. (1995). Methodological principles for structuring an "ontology". In *Proceedings of the IJCAI'95 Workshop on "Basic Ontological Issues in Knowledge Sharing"*, Montréal, Canada.

CHARLET J., BACHIMONT B., MAZUEL L., DHOMBRES F., JAULENT M. & BOUAUD J. (2012a). OntoMenelas : motivation et retour d'expérience sur l'élaboration d'une ontologie noyau de la médecine. *Technique et Science Informatiques*, **31**(1).

CHARLET J., DECLERCK G., DHOMBRES F., GAYET P., MIROUX P. & VANDENBUSSCHE P.-Y. (2012b). Construire une ontologie médicale pour la recherche d'information : problématiques terminologiques et de modélisation. In S. SZULMAN, Ed., *Actes des 23es Journées Ingénierie des Connaissances*, p. 33–48, Paris, France.

GOLBREICH C., GROSJEAN J. & DARMONI S. J. (2013). The Foundational Model of Anatomy in OWL 2 and its use. *Artificial intelligence in medicine*, **57**(2), 119–32.

HORRIDGE M. (2009). *A Practical Guide To Building OWL Ontologies Using Protégé 4 and CO-ODE Tools*. The University Of Manchester, 1.2 edition. 109 pages.

MEJINO J. V., AGONCILLO A. V., RICKARD K. L. & ROSSE C. (2003). Representing complexity in part-whole relationships within the Foundational Model of Anatomy. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, p. 450–4.

REYMONET A. (2007). Modélisation de ressources termino-ontologiques en OWL. In F. TRICHET, Ed., *Actes des 18es Journées Ingénierie des Connaissances*, p. 169–180, Grenoble, France : Cépaduès. ISBN 978.2.85428.790.5.

ROSSE C. & JR J. L. V. M. (2008). The Foundational Model of Anatomy Ontology. In A. BURGER, D. DAVIDSON & R. BALDOCK, Eds., *Anatomy Ontologies for Bioinformatics : Principles and Practice*, chapter 4, p. 59–118. New York : Springer.

ROSSE C. & MEJINO J. L. V. (2003). A reference ontology for biomedical informatics : the Foundational Model of Anatomy. *Journal of biomedical informatics*, **36**(6), 478–500.

VANDENBUSSCHE P.-Y. & CHARLET J. (2009). Méta-modèle général de description de ressources terminologiques et ontologiques. In F. GANDON, Ed., *Actes des 20es Journées Ingénierie des Connaissances*, p. 193–204, Hammamet, Tunisie.

Construction d'une ontologie modulaire pour l'univers de la cuisine numérique

Sylvie DESPRES

LIMICS UMRS 1142, Université Paris 13, Sorbonne Paris Cité, Bobigny, France
sylvie.despres@univ-paris13.fr

Résumé : Dans cet article, nous présentons le cadre méthodologique adopté pour construire l'ontologie modulaire de la cuisine numérique en justifiant le choix de la modularité. Puis nous présentons les différents modèles construits à partir des connaissances acquises auprès des experts du domaine. Nous discutons également les choix effectués pour définir les primitives représentées afin d'être en phase avec le périmètre de l'ontologie. Une fois les modèles de connaissances construits nous présentons notre réflexion sur le choix du langage à utiliser afin d'être au plus près des besoins de raisonnement à produire avec l'ontologie. Nous présentons la manière dont les flux rdf sont engendrés par programmes. Nous abordons également la façon dont les différentes versions de l'ontologie sont gérées.

Mots-clés : modèle de connaissance, ontologie modulaire, ontologie de domaine, enrichissement automatique de modèle ontologique, gestion de versions.

1 Introduction

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 les coups de cœur culinaires, de créer un livre de recettes personnelles en ligne, de trouver des tutoriels vidéos de recettes étapes par étapes, de prendre des cours particuliers de cuisine en ligne, d'organiser les menus à la semaine, etc. Parmi ces applications figurent la tablette QooQ [http://www.qooq.com/tablette/], le projet Le Foodle et sa tablette associée de SEB [http://www.lefoodle.com/] ; les portails de cuisine tels que Cuisine AZ et son application Iphone [http://www.cuisineaz.com/], Cuisinix (cuisine et course) et sa tablette [http://www.cuisinix.fr/] ; l'interface conçue par Chef Jérôme [http://chefjerome.com/search] dont les technologies permettent de relier le monde de la cuisine à celui de la grande distribution de manière automatique et à grande échelle.

Le domaine de la cuisine numérique se caractérise ainsi par le recours aux supports numériques et aux technologies du web pour satisfaire des besoins exprimés dans le domaine de la cuisine portant à la fois sur la réalisation de recettes, la commensalité, la constitution de livre électronique, etc. C'est dans ce contexte que le projet de recherche Open Food System (OFS) [projet-open-food-system] auquel nous participons depuis un an est développé. Il a pour ambition de construire un écosystème de référence permettant de faciliter la préparation des repas grâce à la mise à disposition de contenus, d'appareils et de services innovants. Il vise au développement de solutions pour la cuisine numérique destinées au grand public et adaptées aux différents profils d'utilisateurs. En outre, il a pour objectif de permettre aux amateurs de cuisine des échanges communautaires. Il est également prévu de mettre à la disposition des professionnels et du grand public de nouveaux appareils de cuisson dits intelligents : contrôle automatique des paramètres de cuisson pour un résultat optimal, conservation des qualités organoleptiques et nutritionnelles des aliments cuits.

Une des tâches réalisées par le LIMICS consiste à construire une ontologie pour l'univers de la cuisine numérique. Cette ontologie doit permettre l'élaboration de suggestions nutritionnelles permettant à des internautes de s'alimenter de manière équilibrée, en se faisant plaisir et en permettant de partager leurs expériences culinaires avec des proches. Les suggestions nutritionnelles faites aux utilisateurs sont fondées sur les résultats du Programme National Nutrition Santé (PNNS) (http://www.mangerbouger.fr/pnns) et l'expertise en nutrition de l'Unité de Recherche en Epidémiologie Nutritionnelle (UREN) (http://www.univ-paris13.fr/uren/). Elles prennent en compte les pratiques alimentaires observées dans un échantillon représentatif de familles. Elles comportent en outre des indications sur la saveur de la recette proposée et tiennent compte des préférences exprimées par les utilisateurs de la plateforme OFS. La ressource ontologique construite repose sur les modèles de connaissances des différents domaines représentés.

Dans cet article, nous décrivons le cadre méthodologique adopté pour construire l'ontologie de la cuisine numérique et nous justifions le choix de la modularité pour construire cette ressource. Puis nous présentons les différents modèles construits à partir des connaissances acquises auprès des experts du domaine. La phase d'acquisition est brièvement décrite. Nous discutons également les choix effectués pour définir les primitives représentées afin d'être en phase avec le périmètre de l'ontologie. Une fois les modèles de connaissances construits nous exposons notre réflexion sur le choix du langage à utiliser afin d'être au plus près des besoins de raisonnement à produire avec l'ontologie. Nous présentons la manière dont les flux RDF sont engendrés par programmes. Nous abordons également la façon dont les différentes versions de l'ontologie sont gérées et les limites de l'approche. Puis nous concluons.

2 Cadre méthodologique

Le cadre méthodologique dans lequel nous nous situons est celui du projet NeON (http://www.neon-project.org/). Il s'agit en effet de construire une ontologie modulaire nécessitant d'acquérir de la connaissance auprès des experts des différents domaines impliqués dans le projet, d'examiner la possibilité de réutiliser des ressources existantes et enfin d'évaluer les premiers modules construits.

Chaque module a été élaboré en respectant le cycle classique de construction d'une ontologie (spécification, planification, conceptualisation, formalisation, implémentation) correspondant au scénario 1 de la méthodologie Neon (Suárez-Figueroa, Gómez-Pérez, & Fernández-López, 2012). La spécification des besoins auxquels doit répondre l'ontologie a été décrite dans l'introduction (cf. paragraphe 1). La construction des modèles de connaissances est décrite au paragraphe 3. Les réflexions portent également sur le choix de la version de OWL adopté pour représenter l'ontologie. Une fois les modèles formalisés, l'ontologie est enrichie *via* un traitement en batch permettant de générer des flux RDF à partir de données extérieures. L'implémentation est réalisée par MONDECA, un des partenaires du projet.

Le scénario 2 « Réutilisation et réingénierie de ressources non-ontologiques » a été appliqué à la construction du module concernant les aliments. L'analyse des différentes ressources disponibles a permis de définir un modèle de connaissances relatif aux aliments permettant de répondre aux besoins couverts par l'ontologie. Une restructuration de la ressource sélectionnée a ensuite été réalisée. Cette analyse est décrite au paragraphe 3.

Le scénario 3 « Réutilisation de ressources ontologiques » n'a pas pu être mis en œuvre car s'il existe des ontologies dans le domaine de la nutrition ou de la cuisine numérique, les langues utilisées sont l'anglais ou le portugais [Badra et *al*., 2008 ; Batista et *al*., 2006 ; Cantais, 2005 ; Champin et *al*., 2008 ; Dominguez et *al*., 2006 ; Graca et *al*., 2005 ; Ribeiro et *al*., 2006 ; Snae et *al*., 2008 et 2009 ; Villarias, 2004]. La réutilisation de la ressource ne peut pas être réduite à une traduction des labels figurant dans la ressource. En effet, l'identité culturelle de la cuisine française impacte fortement le modèle des connaissances qui lui sont associées. Les scénarios 4, 5 et 6 dont les objectifs sont centrés sur la réutilisation ne sont par

conséquent pas mis en œuvre. Néanmoins, le travail de recherche des différentes ressources ontologiques a été réalisé en exploitant les moteurs de recherche d'ontologies (swoogle, watson) et une recherche bibliographique classique. Plusieurs projets ont donné lieu à la construction de ressources termino-ontologiques dans le domaine de la cuisine et de la nutrition. Certains de ces projets sont accessibles *via* des URI mais les ressources termino-ontologiques leur correspondant ne sont pas toujours disponibles. Une étude complète de ces différents projets est disponible dans le livrable RTO du projet OFS [Despres, 2013].

Le scénario 8 « Restructuration des ressources ontologiques »est central dans ce travail où nous construisons une ontologie modulaire. Notre approche de modularisation est effectuée par composition. Nous présentons les résultats de cette étape au paragraphe 4.

La construction de chacun des modules est fondée sur des activités d'acquisition de connaissances auprès des experts des domaines concernés. L'acquisition des connaissances est réalisée par les chercheurs en science cognitive de l'Institut Paul Bocuse pour les aspects organoleptiques, les anthropologues participant au projet pour les pratiques alimentaires et par le LIMICS pour les connaissances dans le domaine de la cuisine auprès des chefs et dans le domaine de la nutrition auprès des chercheurs de l'UREN. Deux ateliers d'acquisition des connaissances ont été organisés avec les différents chercheurs dans les domaines impliqués. L'idée était de définir collectivement le périmètre de l'ontologie et expliciter les déterminants associés à chacun des domaines utiles à la construction de l'ontologie. Au cours du premier atelier une séance de brainstorming a été organisée sur les thèmes « recette », « nutrition » et « pratiques alimentaires ». Le second atelier a permis de restituer les connaissances acquises et d'affiner de manière collective le modèle intégrant les différentes perspectives du domaine de la cuisine numérique.

Une première validation des modèles de connaissance a été réalisée par les chefs et l'UREN pour le module aliment et une validation est en cours pour les modules matériel, préparation de base et cuisine. La gestion des versions des modules reste une activité complexe que nous assurons avec Git [http://git-scm.com/], un logiciel de gestion de versions orienté programmation. Chaque module est documenté. L'évaluation de l'apport de l'ontologie à une plateforme n'a pas encore été abordée.

3 Construction d'un modèle de connaissances mettant en jeu plusieurs domaines

L'ontologie doit permettre de raisonner sur des connaissances du domaine de la cuisine numérique afin de produire des suggestions de recettes et de planification de repas répondant aux critères de bien-être défini par le PNNS et procurant du plaisir en respectant les goûts et les habitudes des utilisateurs de la plateforme. Plusieurs domaines de connaissances sont au cœur de ce travail. Nous présentons la trame générale des modèles construits pour chacun d'entre eux et les activités qui ont conduit à leur élaboration (acquisition, réutilisation, etc.).

3.1 Domaine des aliments

Les aliments intervenant dans la réalisation d'une recette sont au cœur du modèle à construire. Ils doivent être considérés selon des points de vue propres à chacun des domaines impliqués dans le projet.

3.1.1 Réutilisation des connaissances

Il existe de nombreuses classifications des aliments construites selon des critères qui ne sont pas forcément pertinents pour notre modèle. 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.

Nous avons finalement travaillé à partir de la classification polyaxiale des aliments utilisée par les chercheurs de l'UREN [Fredot, 2009] qui a été validée par les chefs de l'IBPR.

3.1.2 Modélisation des aliments

La construction du modèle concernant les aliments a été effectuée au cours de deux principales phases. La première a consisté à exploiter les connaissances issues des différentes ressources disponibles concernant la modélisation des aliments. Nous avons exploité la classification polyaxiale de Fredot (cf. *supra*). La seconde phase a consisté à regrouper les différents éléments caractérisant les aliments en sous-groupes en fonction des usages de l'ontologie. Les raisonnements prévus avec l'ontologie pourront permettre d'identifier des aliments en fonction de leurs caractéristiques organoleptique, nutritionnelle, de qualité, de leur état (cru/cuit, fumé, etc.), de leur partie (bulbe, chair/pulpe, gousse, cuisse, suprême, etc.) et de leur origine géographique. Nous envisageons par la suite l'utilisation des « linked data » pour réutiliser le vocabulaire existant pour les caractéristiques géographiques et tenter de faire le lien avec les saisons qui varient d'un pays à l'autre.

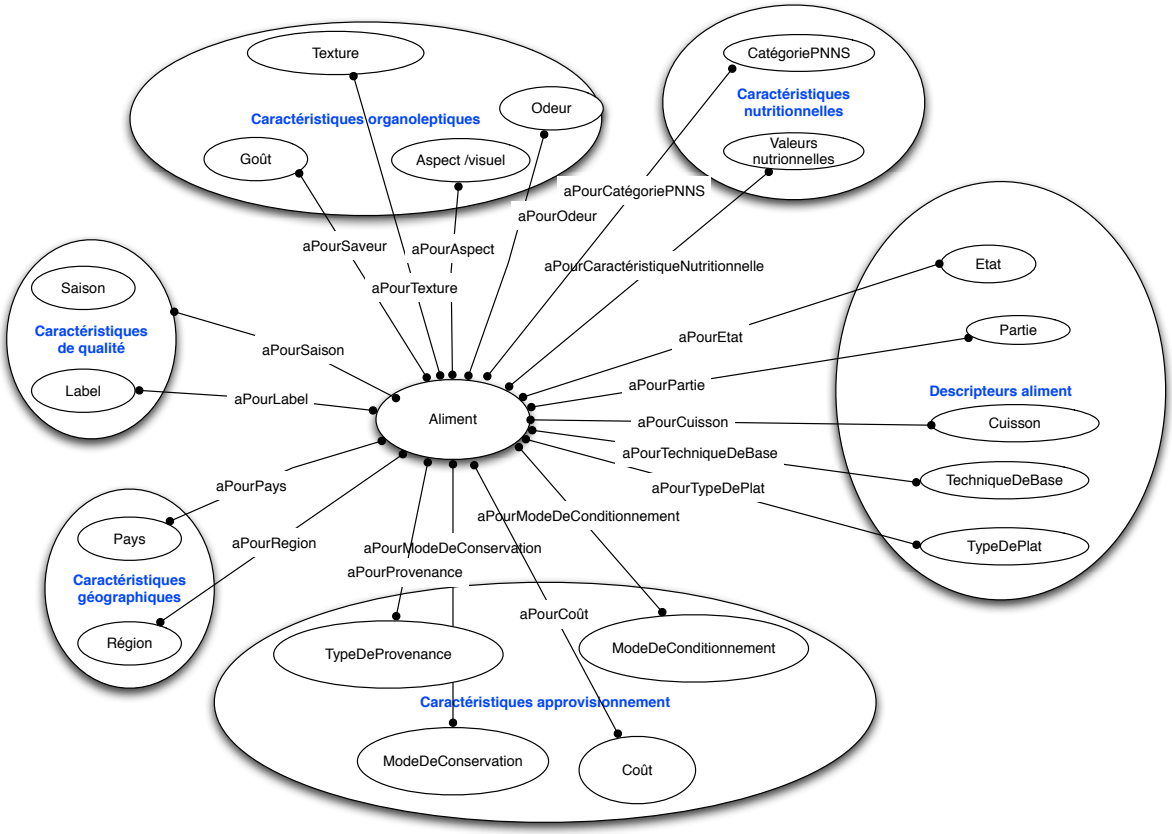


FIGURE 2 – Modèle des aliments

3.2 Domaine de la cuisine

Les connaissances sur le domaine de la cuisine sont essentiellement acquises auprès des chefs « cuisinier » et « pâtissier » du centre de Recherche de l'Institut Paul Bocuse (IPBR). Plusieurs ouvrages papier [Maincent-Morel, 2002], [Charles, 2009], [Deschamps, Deschaintre, 2009], [Chaboissier, Lebigre, 2008] servent également de référence mais ne sont actuellement pas disponibles en version numérique. Nous ne pouvons par conséquent pas exploiter les techniques de construction de ressource terminologique à partir de textes. Nous avons dû avoir recours aux techniques d'acquisition de connaissances auprès des experts du domaine.

3.2.3 Acquisition des connaissances

Les connaissances en jeu concernent les ingrédients, les techniques et préparations de base et le matériel entrant en jeu dans la réalisation d'une recette. Des entretiens ont été réalisés auprès de deux chefs cuisinier et pâtissier. Ils nous ont permis de nous familiariser avec les notions de base en cuisine et en pâtisserie et d'acquérir les connaissances relatives à la modélisation du domaine.

Ces premiers entretiens ont mis en évidence une différence essentielle entre les pratiques dans les deux métiers. En cuisine, des règles de réalisation sont transmises par le chef cuisinier. En pâtisserie les connaissances transmises concernent le plus souvent ce qu'il ne faut pas faire et les préparations peuvent être confectionnées en plus grandes quantités qui seront ensuite conservées au froid.

3.2.4 Modélisation des connaissances

La démarche est identique à la construction du modèle des aliments. Elle s'est déroulée en deux étapes.

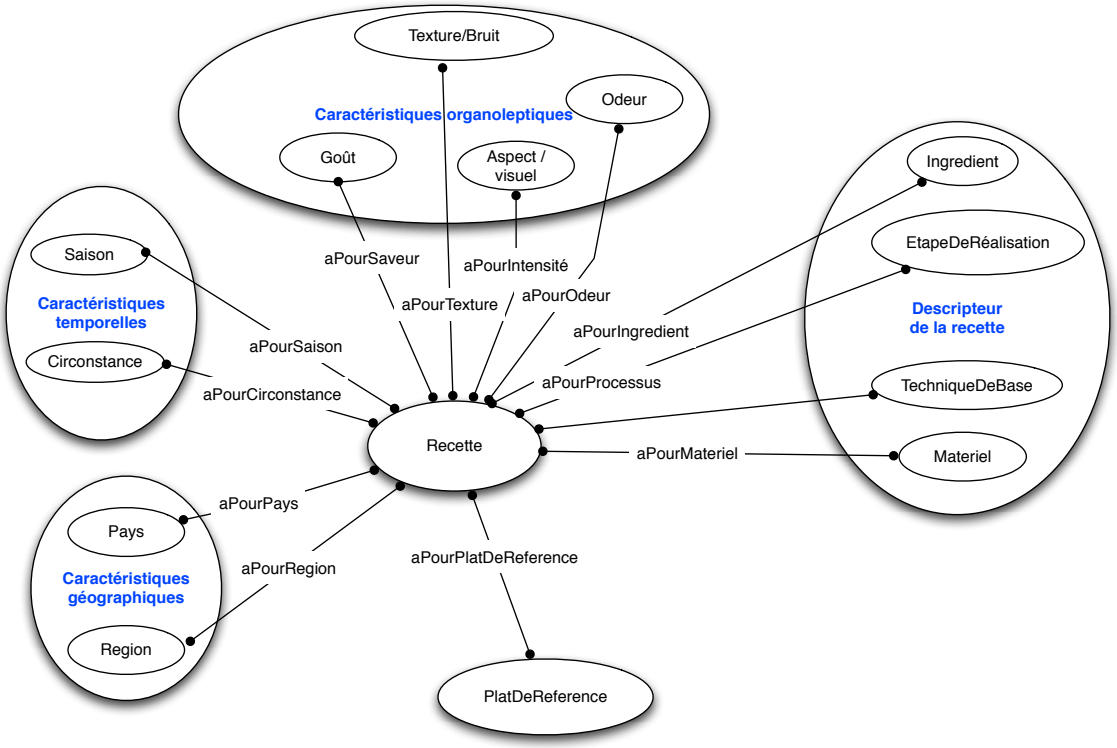


FIGURE 3 – Modèle de la recette

Dans le modèle de la recette, nous avons figuré la notion d'ingrédient. Un ingrédient est simple (tomate) ou composé (pâte feuilletée). Il est constitué d'une unité quantifiée (400g ou 1 cuillère à soupe ou 2 pièces) et d'un aliment lorsqu'il est simple ou d'une préparation de base quand il est composé. Cette préparation de base est elle-même décrite par une recette comportant une liste d'ingrédients. Cette représentation des ingrédients permet l'annotation des recettes de cuisine faisant partie du corpus d'étude (actuellement 55 000 recettes) qui conduit ensuite au classement des recettes en fonction des ingrédients qui les composent.

Un modèle des unités utilisées en cuisine (cuillère à café, verre à moutarde, lichette, etc.) et des unités de mesures internationales (millilitre, gramme) permettant la mise en correspondance entre ces deux systèmes (1 cuillère à soupe de sucre correspond à 10g de sucre) a été élaboré.

3.3 Modèles en cours d'élaboration

L'acquisition des connaissances concernant les domaines du sensoriel et des pratiques alimentaires est actuellement menée par les psychologues cogniticiens auprès des chefs de l'IBP et par des anthropologues auprès des familles participant à l'étude. Nous présentons *infra* les résultats issus des premiers entretiens avec le chef cuisinier et le chef pâtissier de l'IBP et les deux ateliers acquisition des connaissances.

3.3.5 Domaine organoleptique

Le travail de l'équipe du centre de Recherche de l'Institut Paul Bocuse (IPBR) sur les aspects organoleptiques fait intervenir des chercheurs en psychologie cognitive et les chefs. Les connaissances en jeu concernent le goût des aliments, les associations de saveurs lors de la réalisation d'une recette, la caractérisation du goût d'une recette. Comme pour le domaine de la cuisine, les ressources disponibles sont essentiellement sur support papier. Elles comportent les travaux actuels des chercheurs de l'IPBR et plusieurs ouvrages sur les saveurs [Segnit, 2012], [Bassereau & Charvet-Pello, 2011], [Salesse & Gervais, 2012].

Le domaine organoleptique est complexe et les mots pour caractériser les saveurs et les odeurs sont encore peu nombreux et très subjectifs. Des connaissances existent sur les saveurs des aliments simples. L'objectif est de déterminer les éléments organoleptiques permettant la caractérisation des associations des saveurs pour obtenir le rendu d'une recette afin de satisfaire les préférences de l'utilisateur de la plateforme. L'acquisition des connaissances mises en jeu dans ce domaine est en cours de construction et les connaissances acquises seront exploitées dans l'année à venir.

Un atelier a été mis en place pour acquérir les connaissances auprès des psychologues cogniticiens de l'IPB. Une séance de brainstorming a permis d'identifier certains des déterminants concernant les connaissances précédemment énumérées. Les premiers éléments de connaissance liés aux aspects organoleptique font référence aux types de repas (commensalité), à la saveur et au mode de préparation des ingrédients, à la température (glacé, froid, tiède, chaud), à la préférence (hédonisme, émotion) et à la perception (texture en bouche, aspect visuel, saveur, goût, odeur).

3.3.6 Domaine des pratiques alimentaires

Les connaissances dans le domaine des pratiques alimentaires sont acquises par les anthropologues qui observent des familles dans leur quotidien au cours de plusieurs périodes. Ces observations portent sur la façon dont la vie d'une famille est influencée par la nutrition. Elles prennent en compte les aspects individuels des membres de la famille, la réalisation des courses, des repas, etc. Les connaissances en jeu portent sur les questions de : (1) la *mise en route d'une recette ou d'un menu* (- comment l'idée vient : habitude, ingrédients dans le réfrigérateur, discussion, recette internet, recette écrite (imprimée ou livre)) ; - comment se prépare le plat : aller retour entre la recette et les ingrédients (changement de quantité, d'ingrédients...) ; - comment il est mangé : par qui et comment) ; (2) la *gestion des restes* (est-ce que les gens font des restes et pourquoi ?, que font-il des restes quand il y en a ?) ; (3) les *négociations entre les membres de la famille*.

Un atelier a été mis en place pour acquérir les connaissances auprès des anthropologues impliqués dans les familles. Une séance de brainstorming a permis d'identifier certains des déterminants influençant le choix d'une recette ou d'un menu. Ils sont temporels lorsqu'il s'agit des périodes de l'année où des changements d'habitude interviennent (rentrée scolaire - vacances) ou de la saison. Le type des repas est déterminé par un moment (Semaine/WE/Repas de fête) et les convives y participant (individu, famille/avec invité), la composition de la famille, le coût et l'approvisionnement, le contenu du réfrigérateur/congélateur, les restes à accommoder, les matériels disponibles pour la réalisation d'un plat, le temps disponible et la durée de la recette. Le modèle devra par

conséquent prendre en compte ces déterminants pour aboutir à une suggestion satisfaisante pour l'utilisateur de la future plateforme.

3.4 Domaine de la nutrition

Le projet de produire des suggestions orientées « bien-être » nécessite de comprendre les résultats obtenus *via* le PNNS et les déterminants conduisant un individu à choisir une recette.

Les connaissances en jeu sont acquises auprès des experts de l'UREN impliqués dans le projet. Elles portent sur les aliments, la manière appropriée de définir la valeur nutritionnelle d'une recette et la façon de suggérer une planification des repas au cours d'une période de temps pouvant être celle d'une périodicité correspondant à la semaine. Elles s'appuient sur les catégories définies par le PNNS. Les aliments sont classés en grandes catégories et sont listés dans la table Nutrinet qui contient les nutriments caractérisant les aliments intervenant dans les recettes collectées au cours des enquêtes réalisées par l'UREN.

TABLE 1 – *Synthèse des connaissances acquises en nutrition*

Type de suggestion	**Contenu**	**Visualisation**
Qualification de la recette	sel, gras, vitamines, sucre, minéraux, oligo-éléments, nutriments, calories, valeurs nutritionnelles, allergène, intolérance (gluten...), régimes particuliers, "résiste à tout"	Données brutes, échelle, outil de tri
Information nutritionnelle	Données issues de la table de composition des aliments Nutrinet	Indicateur simplifié de correspondance avec les recommandations du PNNS sous la forme d'un curseur graphique
Repas	Vitaminé, fraicheur, basses calories, maintien en forme, prix, saison, bénéfices et apports, conseils, occasions spéciales (fêtes, légumes aux enfants)	Langage iconographique à définir
Familiale (tient compte des contraintes des différents membres de la famille)	Repas de base et variations pour les différents membres, menu de la semaine, calendrier long terme (remise en forme, grossesse, performance), bilan statistique, retour d'information,	Sous forme graphique Alertes pour aider à la prise de conscience
Informations pédagogiques, contextualisées ou non	Explication de la suggestion du point de vue de la nutrition	Menus contextuels
Substitutions d'ingrédients	Une liste d'aliments substituables en fonction du contexte de la recette valide du point de vue de la saveur de la recette et de la nutrition	Sous forme de liste

Les ressources disponibles sont les travaux des chercheurs participant au projet, les documents édités par le PNNS, l'ouvrage sur les aliments et leurs propriétés nutritionnelles fournies par l'UREN [Fredot, 2009], les travaux sur les déterminants réalisés par l'UREN [Olay, 2011], la table de composition des aliments Nutrinet sous forme d'un fichier Excel.

Au cours de l'atelier, un brainstorming a été organisé afin d'identifier le contenu des suggestions, leur nature et leur mode de présentation à l'utilisateur de la plateforme. Elles ont ensuite été regroupées en plusieurs grands types (cf. tableau 1).

3.5 Les acteurs autour du système

Les acteurs du système interrogent le système avec leur propre vocabulaire et leurs propres connaissances dans les domaines de la cuisine, de la nutrition et des saveurs des recettes. Ils expriment également des préférences relatives à leur goût, leurs désirs, leurs aversions et leurs interdits et font partie intégrante de la société. Actuellement, le modèle de la personne est très succinct. Au cours de l'année à venir, il devra prendre en compte les déterminants identifiés auprès des chercheurs dans les domaines de la nutrition, des pratiques alimentaires et des

caractéristiques organoleptiques. Un premier modèle synthétisant les premières connaissances acquises est présenté figure 4.

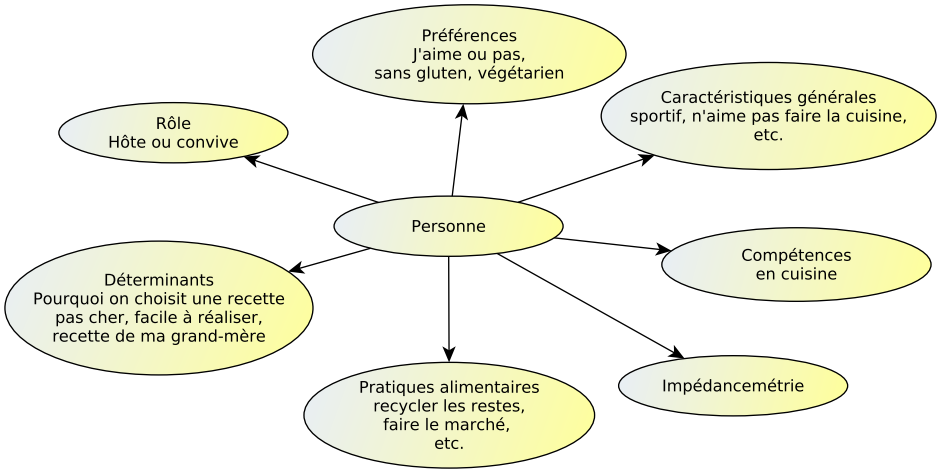


FIGURE 4 – Vers un modèle de l'utilisateur de la plateforme

4 Formalisation

Dans ce paragraphe, nous discutons de la construction modulaire de l'ontologie et du choix du langage pour représenter les modèles associés à chacun des modules.

4.1 Construction d'une ontologie modulaire

Ces dernières années, plusieurs travaux se sont intéressés au développement modulaire des ontologies et à l'échange d'information entre modules ontologiques. Cependant, ces travaux portent essentiellement sur l'intégration d'ontologies existantes en tant que modules dans une ontologie plus large ou sur la gestion d'interrelations entre des ontologies distribuées.

Nous avons fait le choix d'une conceptualisation modulaire dès le début du cycle de développement de l'ontologie. La méthodologie de construction de l'ontologie modulaire adoptée suit une approche par composition. Les différents modules correspondant à chacun des domaines du modèle sont construits et ensuite composés pour constituer l'ontologie globale.

L'existence de liens entre les différents modules et les besoins d'interrogation, de raisonnement et d'inférences - par conséquent, de mise à jour des modules [Stuckenschmidt & Klein, 2003] - nous ont conduit à concevoir une ontologie modulaire (cf. figure 5) qui comportera un module noyau (module ALIMENT) et les modules thématiques suivants : module NUTRITION (concepts spécifiques à la nutrition) ; module CUISINE (relatif à la réalisation des recettes et des liens avec les types de plat) ; module PREPARATION (relatif aux préparations de base associées à une recette) ; module UNITE (relatif aux métriques du domaine de la cuisine (cuillère à café, verre à moutarde) et métriques internationales (gramme, millilitre, etc.) ; module MATERIEL (relatif aux matériels utilisés pour réaliser les recettes) et module ORGANOLEPTIQUE (relatif aux aspects sensoriels caractérisant les aliments et les recettes).

Un module noyau est un module auquel l'ensemble des modules thématiques fait référence. Le point de vue associé à ce module et le vocabulaire utilisé pour caractériser les concepts le constituant sont communs à l'ensemble des modules thématiques y faisant référence. Un module thématique pour un domaine D est une ontologie couvrant un point de vue sur D. Il doit pouvoir être interrogé selon le point de vue qu'il représente

indépendamment des autres modules thématiques. Les différents modules de l'ontologie sont liés par des relations d'interconnexion permettant d'interroger et de raisonner sur l'ontologie globale.

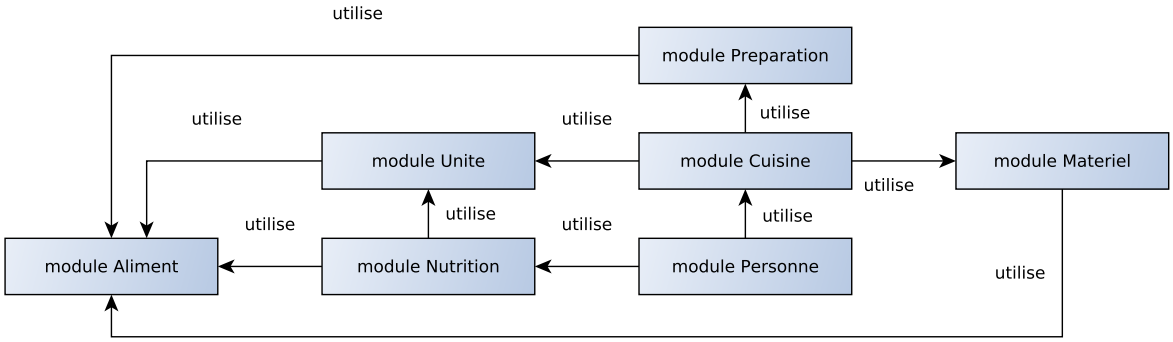


FIGURE 5 – Structuration des modules

4.2 Choix du langage

Une réflexion est actuellement en cours pour décider du langage à utiliser pour construire et raisonner avec l'ontologie.

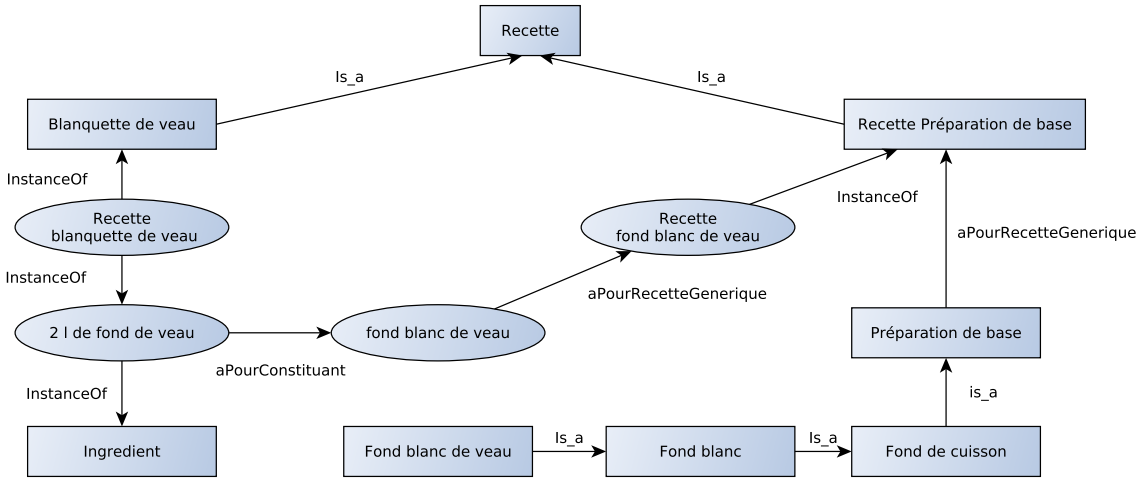
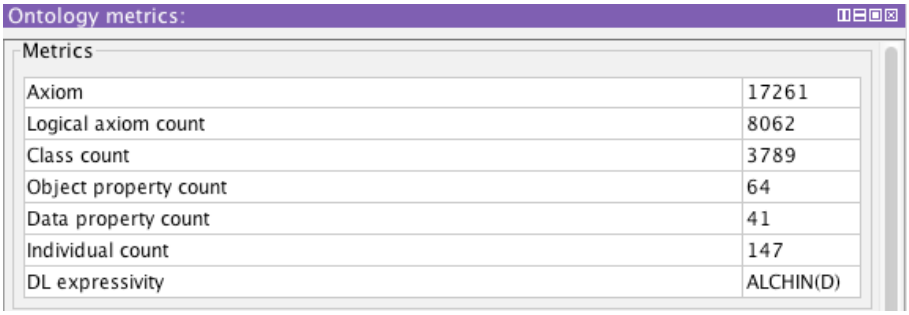


FIGURE 6 – Raisonnement sur la recette

Nous avons en particulier besoin de raisonner sur les classes de l'ontologie pour obtenir des catégories de recettes en référence aux différents modules et de produire des chaînes de raisonnement pour enrichir les recettes par des métadonnées déduites (catégorisation de la recette, variantes de recette, des valeurs nutritionnelles calculées, etc.). Une fois les recettes de base enrichies par la chaîne sémantique, elles sont réintroduites dans le système. L'exemple du traitement des recettes est présenté figure 6. La chaîne de raisonnement liant un plat de référence à une recette générique est également au cœur des réflexions actuelles. Dans ce cas, il est nécessaire de raisonner sur les instances de la recette générique. La notion de « puning » introduite dans OWL2 pourrait constituer une solution envisageable pour optimiser notre modèle. Nous avons dressé une liste de tous les constructeurs OWL apparaissant dans les différents modules et nous nous orientons actuellement vers le choix d'un profil de OWL2 exploitant la partie DL de OWL. La formalisation des interconnexions entre les modules reste à approfondir.

5 Opérationnalisation

L'ontologie comporte actuellement sept modules. La métrique de l'ontologie est présentée figure 7. Les modalités de la publication de l'ontologie sur la toile sont en cours de discussion avec les partenaires du projet. Elle sera en accès libre une fois terminée et publiée.


Ontology metrics:

Metrics

Axiom	17261
Logical axiom count	8062
Class count	3789
Object property count	64
Data property count	41
Individual count	147
DL expressivity	ALCHIN(D)

FIGURE 7 – Métrique de l'ontologie de la cuisine numérique

6 Outils pour la construction et la gestion des versions

La plupart des ajouts et des corrections dans les différents modules sont gérés par programme à l'aide d'un pseudo-langage. Outre s'affranchir des tâches manuelles et répétitives liées à l'utilisation d'un éditeur d'ontologies, l'intérêt de cette technique est de conserver une trace écrite des actions effectuées sur la ressource. Ce pseudo-langage permet de réaliser les opérations répertoriées dans le tableau 2.

Pour des raisons de commodité, nous avons adopté un format tabulaire car il est facile à produire et à modifier à partir d'extraction texte ou de fichiers au format csv (tableur, liste de données fournie par des participants au projet, liste d'autorité fournie par les experts de domaine). Nous avons également implémenté des contrôles syntaxiques automatiques (cf. figure 8) des fichiers produits par l'éditeur d'ontologies Protégé.

```
==> onto03decembre/onto_check.err <==
dest/modulealiment.err:** Missing label for AmiDuChambertin  put ami du chambertin as label
dest/modulealiment.err:** Missing label for BaguetteLaonnaise  put baguette laonnaise as label
dest/modulemateriel.err:** Separators in altlabel for PochePatissiere
dest/modulemateriel.err:** Split it
```

FIGURE 8 – Exemple d'erreurs détectées

La gestion des versions pour les différents intervenants travaillant sur la ressource est assurée par Git un système de contrôle de versions (cvs) classique et éprouvé. Il s'agit d'un système entièrement décentralisé dans lequel chaque auteur peut travailler sur sa version localement et indépendamment du serveur. Une interface cliente multiplateforme GitEye (http://www.collab.net/giteyeapp) lui est associée pour permettre son utilisation par des non informaticiens. Cette approche est loin d'être parfaite car la gestion des différences entre les différentes versions reste difficile à visualiser sous forme de textes. La notion fonctionnelle de OWL est utilisée pour visualiser ces différences. En effet, la sérialisation RDF produite par Protégé n'est pas constante d'une sauvegarde à l'autre. Nous envisageons l'intégration de OWLDIFF pour la gestion des conflits.

TABLE 2 – *Eléments du pseudo langage et exemples associés*

Actions	Cmd	Obj1(Class/Indiv)	Mot-clef	Obj2(Class)	Label | property	AltLabels	Commentaires
Création d'entité (classe ou d'individu)	createClass	CuisseDeCanard	under	ViandeDeCanard	cuisse de canard		27 trouvés dans ingrédients
	createIndiv	SachetDeSel	under	Sachet	sachet de sel		uren
Modification d'entités (classe ou individu), seuls les champs spécifiés sont modifiés	repInClass	Fenouil	with	BulbDeFenouil			
	repInClass	AlimentSansGluten	with	SansGluten			
	repInClass	ViandeDeChevre	with		chèvre	viande de chèvre	chèvre non trouvé dans termino
Déplacement d'entités (classe ou d'individu)	moveClass	VinaigreDeCoing	under	Coing			
	buildTree	NewTree	under	Aliment	from	poisson.thes	
Ajout d'ObjectProperty ou de DataProperty sur des classes ou des individus	propToIndiv	CacDeSel	object	aliment:Sel	aPourAliment		
	propToIndiv	Gramme8	data	8	aPourValeurNumerique		
Constructions syntaxiques particulières (actions sur des ensembles d'entités)	createClass	VarieteCitron	under	Citron	variété_citron		
	moveClass	sub[Citron]	under	VarieteCitron			
	propToClass	aPourTransformation	to	subclasses[VinaigreAromatise]	FabricationVinaigre		
	propToClass	[Ble,Riz,Mais]	object	Poivre	aPourAmi		
	moveClass	["Crevette*"]	under	VarieteCrevette			
	moveClass	sub[Crevette] - [Gambas]	under	VarieteCrevette			
	createClass	[Tomate,Avocat]	under	Chair	chair de $Obj1		
	createClass	sub[LegumeFruit]	under	JusDeFruit	jus de $Obj1		
	propToClass	sub[LegumeFruit]	object	$Obj1	aPourComposant		
	createClass		under	sub[LegumeFruit]	pulpe de $Obj2		
	createClass		under	sub[LegumeFruit]	nectar de $Obj2		
	moveClass	["BulbeDe*"]	under	Bulbe			
	propToClass	["BulbeDe(*)"]	under	$Match1	aPourComposant		
Manipulation d'arborescences (format texte indenté)	exportTree	NewName	under	Coulis	to	coulis.tree	preparation
	importTree		under	PreparationDeBase	from	coulis.tree	

7 Conclusion

Dans cet article, nous avons présenté le cadre méthodologique adopté pour construire l'ontologie modulaire de la cuisine numérique en justifiant le choix de la modularité et les différents modèles construits à la fois à partir des connaissances acquises auprès des experts du domaine et de ressources sur support papier. Le manque de supports numériques ne nous a pas permis la réutilisation de ressources existantes. La ressource ontologique est maintenant suffisamment conséquente pour pouvoir expérimenter à l'échelle les possibilités de raisonnement qu'elle supporte. L'ontologie est en cours d'évaluation *via* un prototype du système qui est implémenté par les partenaires du projet OFS. Les retours de cette expérimentation serviront de base à l'évolution de la ressource.

D'une manière plus générale, les problèmes abordés dans ce travail relèvent de la problématique de la construction d'ontologies modulaires. Ils concernent principalement la définition de ce qu'est un module (sa nature, son périmètre, etc.), la représentation des liens entre les modules, les types de raisonnement associés et la gestion de l'évolution des différents modules. En ce sens ce travail peut servir de base à une réflexion sur le choix d'une construction modulaire dès le début du cycle de développement de l'ontologie. Des critères s'appliquant à toute ontologie mettant en jeu des points de vue pluridisciplinaires pourraient être explicités et ainsi venir affiner ce point dans les méthodologies existantes. Dans ce papier nous n'avons pas abordé le côté collaboratif, une thèse sur ce sujet est en cours.

Références

BADRA, F., BENDAOUD, R., BENTEBITEL, R., CHAMPIN, P-A., COJAN, J., CORDIER, A., DESPRÉS, S., JEAN-DAUBIAS, S., LIEBER, J., MEILENDER, T., MILLE, A., NAUER, E., NAPOLI, A., TOUSSAINT, Y. (2008) Taaable: Text Mining, Ontology Engineering, and Hierarchical Classification for Textual Case-Based Cooking. In Computer Cooking Contest - Workshop at European Conference on Case-Based Reasoning (ECCBR'08), Schaaf, M. ed. Trier, Germany. pp. 219-228.

BASSEREAU J.F. & CHARVET-PELLO R. (2011). Dictionnaire des mots du sensoriel. Editions TEC & DOC Lavoisier. 519p.

BATISTA, F., MAMEDE, N.J., PARDAL, J.P., RIBEIRO, R., VAZ, P. (2006) Ontology construction: cooking domain. « Technical Report, INESC-ID. Lisbon.

CANTAIS, J., DOMINGUEZ, D., GIGANTE, V., LAERA, L., TAMMA,V. (2005). An example of food ontology for diabetes control. In C. Welty and A. Gangemi, "Working notes of the ISWC 2005 workshop on Ontology Patterns for the Semantic Web", Galway, Ireland, 2005.11.07.

CHABOISSIER, D. & LEBIGRE, D. (2008). Compagnon et Maitre pâtissier – Tome1, 2e édition. Editeur : Villette (Jérôme), 198p.

CHAMPIN, P-A., CORDIER, A., DESPRÉS, S., FUCHS, B., LIEBER, J., MILLE, A. (2008). Construction manuelle de la partie haute d'une ontologie modulaire destinée à une annotation de cas textuels - étude de cas pour une application culinaire dans le cadre du projet Taaable. In 16ème atelier de Raisonnement à Partir de Cas, Nancy.

CHARLES, G. (2009). La cuisine expliquée. Editions BPI. 735p.

DESCHAMPS, B. & DESCHAINTRE, J.C. (2009). Le livre du pâtissier. Editions LT Jacques Lanore. 368p.

DESPRES, S. (2013). RTO en Nutrition. Livrable FL4.1.01 2013 09 30 du projet OFS, 36p.

DOMINGUEZ, D., GRASSO, F., MILLER, T., SERAFIN, R. (2006) PIPS: An Integrated Environment for Health Care Delivery and Healthy Lifestyle Support, ECAI 2006.

FREDOT, E. (2012). Connaissances des aliments. Base alimentaires et nutritionnelles de la diététique. 3ème édition. Editions TEC et DOC. Lavoisier. 613p.

GRAÇA, J., MOURÃO M., ANUNCIAÇÃO, O., MONTEIRO, P., PINTO, H. S., LOUREIRO, V. (2005) « Ontology building process: The wine domain ». In Proceedings of EFIT 2005.

MAINCENT-MOREL, M. La cuisine de référence. Techniques et préparations de base. Fiches techniques de fabrication. Editions BPI. 1040p.

OLAY, A. (2011). Les déterminants des choix des plats et des recettes de cuisine. Mémoire de stage. Master des sciences du sport. Université Paris Descartes.

RIBEIRO, R., BATISTA, F., PARDAL, J.P., MAMEDE, N.J., PINTO, H.F. (2006) Cooking an Ontology. In 12th International Conference on AI : Methodology, Systems, Applications. Berlin, pp.213- 221.

SNAE, C., BRUECKNER, M. (2009) «Personal Health Assistance Service Expert System (PHASES) », International Journal of Biological and Life Sciences, 4-2.

SNAE C., BRÜCKNER, M. (2008) FOODS: A Food-Oriented Ontology-Driven System In Second IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2008)

SEGNIT N. (2012) - Le répertoire des saveurs, Marabout, 493p.

SALESSE, R. & GERVAIS, R. (2012). Odorat et goût. de la neurologie des sens chimiques aux applications. 539p.

STUCKENSCHMIDT, H. PARENT, C. & SPACCAPIETRA, S. éditeurs (2009). Modular Ontologies - Concepts, Theories and Techniques for Knowledge Modularization. Springer, 378p.

SUAREZ-FIGUEROA, M. C., GOMEZ-PEREZ, A., & FERNANDEZ-LOPEZ, M. (2012). The NeOn Methodology for Ontology Engineering. In M. C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, e& A. Gangemi (Eds.), *Ontology Engineering in a Networked World* (pp. 9–34). Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-642-24794-1_2.

VILLARIAS, L.G. (2004) « Ontology-based semantic querying of the Web with respect to food recipes», Informatics and Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark. Master Thesis. ISSN 1601-233X, 2004.

IDOSCHISTO : une extension de l'ontologie noyau des maladies infectieuses (IDO-Core) pour la schistosomiase

Gaoussou Camara[1,2,3], Sylvie Despres[1], Moussa Lo[2]

[1] INSERM, U1142, LIMICS, F-75006, Paris, France.
Université Paris 13, Sorbonne Paris Cité, LIMICS, (UMR_S 1142), F-93430, Villetaneuse, France.
sylvie.despres@univ-paris13.fr
[2] LANI, Université Gaston Berger, B.P. 234 Saint-Louis, Sénégal.
moussa.lo@ugb.edu.sn
[3] Université Alioune Diop de Bambey, B.P. 30, Bambey, Sénégal.
gaoussou.camara@uadb.edu.sn

Résumé : Cet article décrit la construction de l'ontologie de la schistosomiase (IDOSCHISTO). Elle est conçue comme une extension de l'ontologie noyau des maladies infectieuses (IDO) afin d'assurer son interopérabilité avec les ontologies de domaine des maladies infectieuses existantes. IDOSCHISTO importe intégralement ou partiellement des ontologies du domaine médical conçues pour des besoins spécifiques propres aux maladies infectieuses. Il s'agit par exemple des ontologies concernant la transmission des agents pathogènes, le diagnostic, les études cliniques et biologiques, etc. IDOSCHISTO est une ontologie modulaire de la schistosomiase intégrant les points de vue épidémiologique, clinique et biologique de la maladie. Elle a pour finalité de devenir une référence dans le domaine des maladies infectieuses en facilitant la communication et l'interopérabilité entre les différents acteurs impliqués et en supportant le raisonnement dans des contextes d'usage variés tels que la veille sanitaire, l'aide au diagnostic et à la prescription médicamenteuse, la gestion du dossier patient, les annotations biologiques, etc.

Mots-clés : ontologie de domaine, ontologie noyau, ontologie modulaire, maladie infectieuse, schistosomiase.

1. Introduction

La dynamique des populations humaines et leurs interactions avec l'environnement favorisent la propagation rapide des maladies infectieuses. La maîtrise du mode de contamination de la maladie et de ses facteurs de risque est indispensable pour en prévenir l'émergence ou en contrôler l'évolution. Dans le cas de la schistosomiase, une maladie parasitaire dont le parasite vit dans les eaux douces, les interactions des populations avec les points d'eau augmentent le risque de propagation. Ainsi, il devient indispensable de mettre en place un système de veille épidémiologique pour la prévention et le contrôle de ces types de phénomènes. L'analyse des risques et la prise de décision, phases incontournables du processus de veille (Camara et al., 2012a), nécessitent l'implication de plusieurs acteurs aux profils différents et ayant chacun une vision partielle du phénomène étudié. Leur appartenance à des communautés différentes fait qu'ils n'utilisent pas les mêmes vocabulaires pour désigner les mêmes concepts du domaine. Par exemple, un médecin parlera de *patient* alors que l'épidémiologiste parlera d'*hôte infecté*. Pour la veille d'une maladie spécifique telle que la schistosomiase, il est par conséquent nécessaire de disposer d'une ontologie jouant le rôle de médiateur entre les différents acteurs mais aussi d'un modèle formel des connaissances du domaine pour automatiser les raisonnements dans un contexte de veille épidémiologique.

Des ontologies relatives à différentes maladies ont été construites pour améliorer les études biologiques, les traitements cliniques des patients, etc., comme par exemple dans (Lin et al., 2011). Néanmoins, ces ontologies ne sont pas conçues pour la veille et la prévention en santé publique d'une maladie infectieuse. En outre, il n'existe aucune ontologie décrivant la schistosomiase.

Dans cet article, nous proposons la construction d'une ontologie de domaine de la schistosomiase (IDOSCHISTO[1]). . IDOSCHISTO est conçue en tenant compte des différentes perspectives sur la maladie (biologique, clinique et épidémiologique) dont l'intégration est indispensable pour le contrôle et la prévention de sa propagation. L'intérêt visé par cette décomposition modulaire est également de faciliter son utilisation partielle dans ces sous-domaines respectifs. La méthodologie de construction est fondée sur la réutilisation d'ontologies noyau du domaine des maladies infectieuses et d'une ontologie de fondement. Des ontologies spécifiques de domaine existantes sont réutilisées au niveau de la couche spécifique de domaine suivant la recommandation d'OBO Foundry (Smith et al., 2007).

Dans la suite de cet article, nous présentons en section 2 le cadre méthodologique de construction de l'ontologie. La modularisation de l'ontologie est détaillée en section 3. La section 4 décrit les choix de réutilisation adoptés. La section 5 présente la formalisation et la métrique d'IDOSCHISTO. Un cas d'usage de cette ontologie pour l'annotation des données de la schistosomiase de la localité de Richard Toll au Sénégal est fourni dans la section 6. Les résultats de l'évaluation de l'ontologie sont présentés en section 7. Nous concluons et discutons les résultats en section 8.

2. Méthodologie de construction de l'ontologie de la schistosomiase

La construction de l'ontologie de domaine de la schistosomiase suit les différentes étapes définies dans la méthodologie NeON (Suárez-Figueroa et al., 2012). Nous avons recours à plusieurs des scénarios de cette méthodologie pour construire l'ontologie de la schistosomiase.

1 scénario 1 : de la spécification à l'implémentation
2 scénario 3 : la réutilisation de ressources ontologiques
3 scénario 5 : la réutilisation et la fusion de ressources ontologiques
4 scénario 7 : la réutilisation de patron de conception d'ontologie

Nous avons mis en œuvre ces scénarios en respectant les étapes définies dans le *scénario 1* : *spécification, acquisition, modélisation, formalisation, implémentation* et *évaluation*.

1 La *spécification* : la construction de l'ontologie est guidée par l'objectif de support à la veille de la propagation de la schistosomiase. La finalité d'IDOSCHISTO est l'annotation des ressources de la veille, le travail collaboratif entre les acteurs et les organisations impliqués dans la veille (communication, interopérabilité des données et des applications), etc. A cet objectif, s'ajoute le besoin d'assister les activités cliniques et biologiques.
2 L'*acquisition* : cette tâche est réalisée *via* des entretiens avec les experts[2] des domaines respectifs modélisés et l'exploitation de ressources non-ontologiques (rapports, articles, etc.) et ontologiques (principalement le portail « OBO Foundry » des ontologies du domaine biomédical).
3 La *modélisation* : une première approche a consisté à réaliser des modèles conceptuels en vue de la structuration des connaissances et serviront de support à l'évaluation menée avec les experts du domaine. Des diagrammes de classes UML sont utilisés

[1] https://github.com/gaoussoucamara/idoschisto/blob/master/idoschisto.owl

[2] Les experts impliqués dans la construction de l'ontologie sont principalement ceux du Programme National de Lutte contre la Schistosomiase au Sénégal, l'ONG Espoir Pour la Santé de Saint-Louis du Sénégal et une biologiste spécialisée dans la schistosomiase.

pour cette modélisation et des représentations tabulaires permettant une visualisation plus synthétique de certaines connaissances sont proposées.

4 La *formalisation et l'implémentation* : le niveau logique requis pour faire des raisonnements sur l'ontologie est la logique du premier ordre. Les objectifs de nos travaux et les besoins qui leurs sont associés nous orientent vers le choix du langage OWL-DL. Ce langage inclut toute la sémantique formelle des logiques de description (DL) et les capacités de raisonnement qui en découlent, dans un standard de représentation d'ontologie. Les services d'inférences offerts par ce langage, nous permettront de fournir pour le système de veille, des possibilités de raisonnement variées telles que la déduction de connaissances implicites à partir de connaissances représentées explicitement, la vérification de la cohérence de modèles, la classification d'instances, etc. L'utilisation de l'éditeur Protégé qui produit la représentation formelle et propose une sauvegarde du modèle dans un langage de représentation d'ontologie exploitable par la machine nous affranchit de l'étape d'implémentation.

5 L'*évaluation* : (1) La première évaluation est intervenue au cours de la conceptualisation en faisant valider le contenu des ontologies par les experts. Ce processus a été réalisé avec les experts en épidémiologie et de la schistosomiase. Au cours des entretiens, les modèles construits ont été présentés. La validité et la cohérence des connaissances représentées ont été discutées. Les omissions susceptibles d'améliorer la couverture du domaine en rapport avec les objectifs de l'ontologie ont été identifiées. (2) Au cours de la deuxième phase de l'évaluation, nous avons procédé à des tests de consistance en utilisant les raisonneurs inclus dans l'éditeur Protégé. Outre, ce test d'incohérence de classification, nous faisons une classification automatique de l'ontologie en intégrant les nouvelles classifications inférées. OOPS (Poveda-Villalón et al., 2012) a également été utilisé pour déceler des incohérences ou des redondances afin de les corriger (automatiquement ou manuellement). (3) La troisième étape de la validation est liée à l'implémentation des ontologies dans une application réelle afin de tester leur opérationnalité et mesurer leur apport au domaine d'étude considéré. Elle nécessite un temps de développement important dont nous ne disposions pas. Elle a, dans un premier temps, été remplacée par l'utilisation de l'interface d'interrogation de Protégé pour exécuter des requêtes SPARQL sur l'ontologie ou la base d'annotations.

3. Modularisation de l'ontologie du domaine de la schistosomiase

La première étape a consisté à construire l'ontologie IDOSCHISTO selon un modèle en couches d'abstraction permettant ainsi la réutilisation d'ontologies noyau et d'une ontologie de fondement. Ensuite, nous avons pris en compte les points de vue épidémiologique, clinique et biologique sur la maladie au niveau de la couche spécifique. Chacun de ces points de vue a été modélisé comme un module ontologique. Enfin, ces différentes perspectives ont été intégrées dans une ontologie unique en modélisant les *relations inter-perspectives*, c'est-à-dire les relations entre les concepts de perspectives différentes. Les relations entre les concepts de même perspective sont appelées ici les *relations intra-perspective* (Camara et al., 2013).

3.1. Modélisation en niveau d'abstraction de chaque module

Le cadre conceptuel pour construire l'ontologie de la schistosomiase est structuré en trois couches : l'ontologie de fondement (Falbo et al., 2010), l'ontologie noyau (Scherp et al., 2011) et les ontologies spécifiques. Les ontologies spécifiques consistent ici à modéliser les perspectives épidémiologique, clinique et biologique.

La couche noyau concerne le domaine des maladies infectieuses de manière générale. Une unique ontologie noyau des maladies infectieuses, appelée *Infectious Disease Ontology* (IDO-Core), existe dans le domaine. IDO-Core couvre les concepts généraux du domaine des maladies infectieuses (agent pathogène, gène, cellule, organe, organisme, population, hôte, vecteur, humain, etc.) et leurs relations. IDO-Core est reliée à l'ontologie de fondement BFO. Les entités modélisées dans IDO-Core relèvent de plusieurs perspectives : biologique (propriétés biologiques des agents pathogènes et leurs interactions avec l'organisme des individus infectés), clinique (symptôme, diagnostic, traitement, etc.) et épidémiologique (étude des mécanismes de propagation des maladies infectieuses dans la population d'individus et leurs moyens de lutte et de prévention). Ainsi, IDO-Core s'avère pertinente pour constituer le noyau des trois modules correspondants aux différentes perspectives définies. En outre, sa réutilisation permet de garantir une interopérabilité et une réutilisabilité dans le domaine des maladies infectieuses. Néanmoins, IDO-Core ne prend pas bien en compte la perspective épidémiologique traitant du mode de propagation de la maladie et les stratégies de contrôle et de prévention. Même si plusieurs concepts liés au sous-domaine épidémiologique sont présents dans cette ontologie noyau, ils ne sont pas mis en relation pour refléter le mécanisme de la propagation des maladies infectieuse. C'est ainsi qu'IDO-Core est utilisée conjointement avec l'ontologie noyau de la propagation des maladies infectieuses (IDSDO-Core[3]) que nous avons proposée et présentée dans (Camara et al., 2014, 2012b). Ainsi, elles constituent toutes les deux un noyau complet pour les modules spécifiques de la schistosomiase.

Le choix de l'ontologie de fondement à réutiliser est fondé sur trois critères : (i) la cohérence de la catégorisation des concepts de processus, événement, état et objet vis-à-vis de leurs sémantiques dans le domaine des maladies infectieuses, (ii) la consistance de la réutilisation des relations entre ces concepts pour couvrir les relations spécifiques à notre domaine et (iii) la réutilisation d'une ontologie de fondement par les ontologies noyau du domaine des maladies infectieuses, c'est-à-dire IDO-Core et IDSDO-Core.

Enfin, un élément de choix guidé par le besoin de maintenir notre objectif de réutilisabilité et d'interopérabilité dans le domaine des maladies infectieuses, nous a amené à choisir BFO comme ontologie de fondement. L'ontologie de fondement BFO fournit des concepts et des relations pertinents pour la construction d'ontologie de domaine en biomédecine (Grenon et al., 2004). Elle est déjà réutilisée par IDO-Core et IDSDO-Core. Ces différentes caractéristiques la rendent plus appropriée pour la modélisation ontologique du domaine des maladies infectieuses que d'autres ontologies de fondement existantes.

3.2. Modélisation multi-perspectives modulaires

Plusieurs points de vue sur les maladies, que nous appelons désormais perspectives, sont adoptés en médecine en fonction des études réalisées :

1 *Perspective biologique* : la biologie en tant que discipline couvre un domaine très large allant de la molécule à l'écosystème en passant par la cellule, l'organe et la population. Dans ce travail, nous nous limiterons à l'étude de l'interaction biologique de l'agent pathogène avec l'organisme, la réaction physiopathologie des hôtes à la maladie, la taxinomie et le cycle de vie des êtres vivants. L'objectif est de fournir une terminologie permettant d'annoter les données et les ressources liées à ces aspects biologiques de la schistosomiase. Cette même ressource ontologique est également réutilisable par les spécialistes biologistes de la schistosomiase dans leurs recherches scientifiques respectives.

2 *Perspective clinique* : les activités cliniques concernent essentiellement le traitement des patients. La perspective clinique de l'ontologie fournit une spécification des connaissances liées aux activités cliniques comme le diagnostic (les signes et

[3] https://github.com/gaoussoucamara/idoschisto/blob/master/idsdo-core.owl

symptômes chez le patient) et son mode de traitement (prescription médicamenteuse). Cette partie de l'ontologie devrait faciliter l'intégration des données médicales sur la schistosomiase dans les systèmes d'information hospitaliers, et améliorer la prise en charge des patients souffrant de cette maladie.

3 *Perspective épidémiologique* : elle concerne principalement les épidémiologistes et est celle des systèmes de veille en santé publique. Le rôle de l'épidémiologiste est d'étudier les différents facteurs intervenant dans l'apparition et la distribution d'une maladie et les moyens à mettre en œuvre pour sa prévention et son contrôle. La veille épidémiologique est alors mise en œuvre à la fois, pour surveiller les facteurs de risque de propagation d'une maladie, pour analyser l'impact des événements détectés et pour suggérer des plans d'action en cas de risque identifié. Cette perspective de l'ontologie est donc principalement conçue pour représenter les connaissances sur l'épidémiologie de la maladie et servir à la prévention et au contrôle de la maladie.

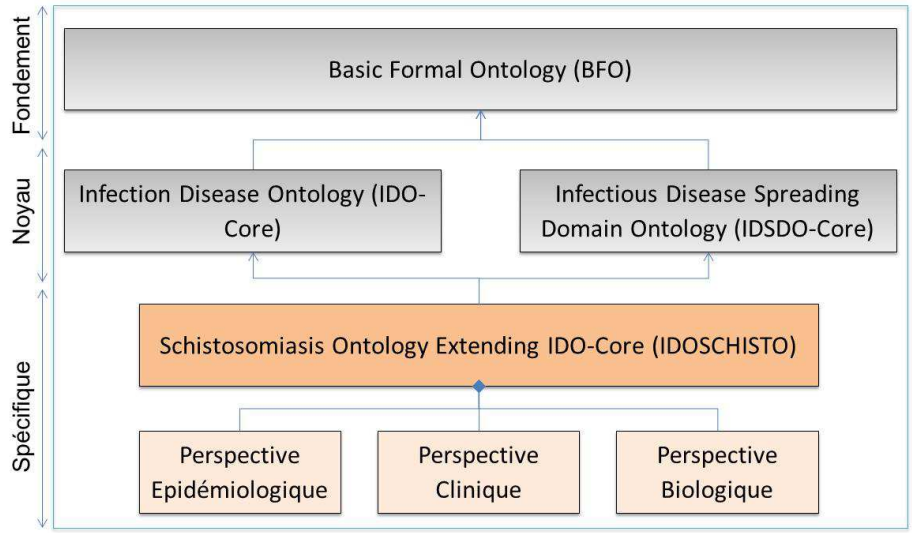


Figure 1 : Modularisation de l'ontologie de domaine de la schistosomiase (IDOSCHISTO)

Outre la cohérence dans la modélisation, la structure modulaire envisagée vise surtout la possibilité d'une utilisation partielle (Doran et al., 2007) de l'ontologie dans les disciplines correspondant à ces perspectives. Cette utilisation partielle permet de faire abstraction des concepts et relations non pertinents pour le sous-domaine considéré mais aussi de manipuler une ontologie de plus petite taille, ce qui renforce sans doute la performance de l'application au moins sur le plan de la complexité (chargement de l'ontologie, rapidité de l'exécution des requêtes, etc.).

3.3. Modélisation des relations inter-perspectives

Bien que la veille relève de la perspective épidémiologique, elle s'étend à toutes les perspectives dont la prise en compte est indispensable pour une bonne analyse des risques mais aussi des décisions adéquates. Ces différentes perspectives ne sont pas totalement indépendantes et sont plutôt fortement liées. Par exemple, ce sont les études biologiques qui permettent de produire un médicament contre la maladie. Lequel médicament est prescrit au patient lorsqu'il est diagnostiqué malade de la schistosomiase. Le traitement des malades est une des stratégies de contrôle de la propagation des maladies transmissibles comme c'est le cas pour la schistosomiase. Ainsi, l'intégration de ces différentes perspectives passe par la modélisation des relations existantes entre leurs concepts. Dans le domaine de la médecine il n'existe à l'heure actuelle qu'une seule ontologie permettant de définir les relations entre les concepts de ce domaine. Il s'agit de Relation Ontology (RO) (Smith et al., 2005). Par ailleurs,

une autre ontologie appelée RO-Bridge a été conçue ajoutant des contraintes de domaine (*domain*) et de co-domaine (*range*) aux relations définies dans RO pour les concepts de BFO. Ces deux ontologies sont donc utilisées et enrichies pour modéliser les relations dans les différents modules de l'ontologie. Il faut noter RO contient uniquement des relations et son importation ajoute dans IDOSCHISTO les relations qu'elle définit. Et puisque IDOSCHISTO a déjà importé BFO, l'importation de RO-Bridge permet d'ajouter les contraintes de domaine et de co-domaine entre RO et BFO.

4. La réutilisation de ressources ontologiques existantes

La réutilisation de ressources ontologiques existantes est un des scénarios proposés par la méthodologie NeOn. Elle est aussi une recommandation d'OBO Foundry pour les ontologies construites dans le domaine biomédical et aspirant à être publiées dans son portail[4]. La réutilisation d'ontologies de fondement et noyau constitue déjà une adhésion à ce principe visant essentiellement une interopérabilité des ontologies construites dans le domaine biomédical. Pour cela, deux procédés sont possibles : (i) la réutilisation complète ou partielle par importation dès le début de la construction ou (ii) l'alignement avec les ontologies existantes après construction. Notre approche combine les deux procédés. Nous faisons l'importation complète des ontologies ayant un niveau de développement mature dont le contenu global est pertinent. C'est le cas de BFO, d'IDO-Core et d'IDSDO-Core. Pour les autres ontologies stables mais dont une partie seulement nous intéresse, nous utilisons l'outil Ontofox[5] (Xiang et al., 2010) pour extraire la partie en question. C'est par exemple le cas de l'extraction de la classification des schistosomes, parasite de la schistosomiase, depuis la taxinomie NCBI. Pour les ontologies utiles aux différents sous-modules et qui ne sont pas encore stables ou validées, nous proposons de procéder à un alignement de nos ontologies une fois leur construction terminée.

Tableau 1 : Extrait des ontologies réutilisées dans la construction d'IDOSCHISTO

Ontologies	Epi.	Clin.	Bio.
Basic Formal Ontology (BFO)	✓	✓	✓
Infectious Disease Ontology (IDO-Core)	✓	✓	✓
Information Artifact Ontology (IAO)	✓	✓	✓
Infectious Disease Spreading Ontology (IDSDO-Core)	✓	×	×
Pathogen transmission (PT)	✓	×	×
Human Disease Ontology (DOID)	×	✓	×
Ontology for General Medical Science (OGMS)	×	✓	×
Symptom Ontology (SO)	×	✓	×
Vaccine Ontology (VO)	×	✓	×
Ontology of Biomedical Investigation (OBI)	×	✓	✓
Chemical Entities of Biological Interest (ChEBI)	×	×	✓
Protein Ontology (PO)	×	×	✓
NCBI Taxonomy Database	×	×	✓

Les ressources ontologiques répertoriées (cf. Tableau 1) ont été étudiées manuellement avec l'éditeur Protégé. Cette étude a consisté à rechercher la présence ou à déterminer la couverture des entités des sous-domaines que nous avons identifiées et modélisés dans la phase conceptuelle. Ensuite, les classes (ou relations) pertinentes pour nos ontologies sont

[4] http://www.obofoundry.org/

[5] http://ontofox.hegroup.org/

sélectionnées pour leur extraction avec Ontofox. L'ontologie en question peut être sélectionnée dans une liste déroulante déjà fournie ou être renseignée à travers son URI. Pour une entité donnée de l'ontologie, Ontofox permet d'extraire les *sous-classes* et la *hiérarchie* jusqu'à une *superclasse* qu'on aura indiquée. Les axiomes, relations et annotations associées à la classe peuvent aussi être extraites.

En plus d'IDO-Core, d'IDSDO-Core et de BFO, il existe plusieurs ontologies de domaine représentant les connaissances générales (indépendantes d'une maladie donnée) liées aux domaines des protéines, de la taxonomie des êtres vivants, des signes et symptômes, du diagnostic, du traitement, des médicaments, de la transmission de pathogène, etc. Ces ontologies offrent déjà des modèles cohérents et génériques que nous pouvons importer (partiellement ou intégralement) dans IDOSCHISTO. Étant donné qu'IDOSCHISTO est décomposée en perspectives, il nous a d'abord fallu déterminer, pour chacune de ces ontologies à importer, quelle perspective était pertinente ou compatible. Trois critères ont été établis pour déterminer la compatibilité de ces ontologies avec les perspectives retenues :

1 *critère épidémiologique* : contamination, transmission, propagation, facteur de risque, prévention, contrôle, etc.
2 *critère clinique* : signe, symptôme, diagnostic, traitement, médicament, etc.
3 *critère biologique* : gène, protéine, physiopathologie, cycle de vie, etc.

Dans le Tableau 1, en plus d'IDO-Core, d'IDSDO-Core et de BFO que nous importons intégralement, nous fournissons la liste des ontologies de domaine réutilisées partiellement ou intégralement pour chaque perspective. Il faut noter qu'une ontologie de domaine peut être compatible avec plusieurs perspectives. Le symbole ✓ signifie que cette ontologie de domaine est compatible avec cette perspective et le symbole ✗ signifie qu'elle est incompatible ou juste non pertinente pour cette perspective.

5. Formalisation de l'ontologie IDOSCHISTO

L'ontologie de la schistosomiase est ainsi construite en trois étapes :

1 L'importation intégrale ou partielle d'un ensemble d'ontologies existantes et dont les états de développement sont suffisamment avancés. BFO, IDO-Core, IDSDO-Core et RO ont été intégralement réutilisées. Les autres ontologies partiellement réutilisées sont listées dans le Tableau 1. Par exemple, les différentes espèces de parasite schistosome causant la schistosomiase ont été importées de la taxonomie NCBI et leurs cycles de vie ont été extraits d'Ontology for Parasite LifeCycle (OPL). Le choix de cette dernière ontologie est guidé par le besoin de superviser la mutation des parasites schistosome observée récemment par les experts biologistes et conduisant à de nouveaux types de schistosomiase.
2 L'ajout des concepts, relations et types de données propres à la schistosomiase et n'existant pas dans les ontologies importées. Ces éléments sont obtenus à partir des connaissances acquises lors des entretiens auprès des experts et l'exploitation des ressources non ontologiques tels que les publications scientifiques, les rapports d'étude, etc.
3 L'alignement manuel des concepts, relations et types de données ajoutés avec des ressources termino-ontologiques existantes que nous n'avons pas importées. C'est le cas par exemple des ontologies sur les symptômes (Symptom Ontology), les médicaments (Drug Ontology), le vaccin (Vaccine Ontology), Ontology for General Medical Science, etc.

En résumé, les ontologies importées ne couvraient pas tous les concepts de la schistosomiase. Nous avons ajouté une soixantaine de concepts centraux de la schistosomiase. Notons que ces ajouts tiennent compte de la distinction abstraite dans BFO entre les *continuants* et les *occurrents* (Simons and Melia, 2000). Les continuants représentent les

entités sans partie temporelle tels que les *objets* alors que les occurrents représentent la classe des entités dynamiques tels que les *processus*.

Parmi les continuants, nous avons par exemple les modes de diagnostic, des conditions environnementales favorisant ou non la présence de la schistosomiase, etc. Nous n'avons pas directement rattachés ces concepts aux « Continuants » mais nous les avons répartis dans des sous-classes en fonction de leur cohérence. Les processus inhérents au domaine de la schistosomiase sont liés aux stratégies de contrôle et de prévention que nous avons ajoutées après exploitation des connaissances acquises auprès des experts et des ressources documentaires. Nous pouvons aussi citer les principales activités des populations les mettant en contact avec les points d'eau (« water_body »).

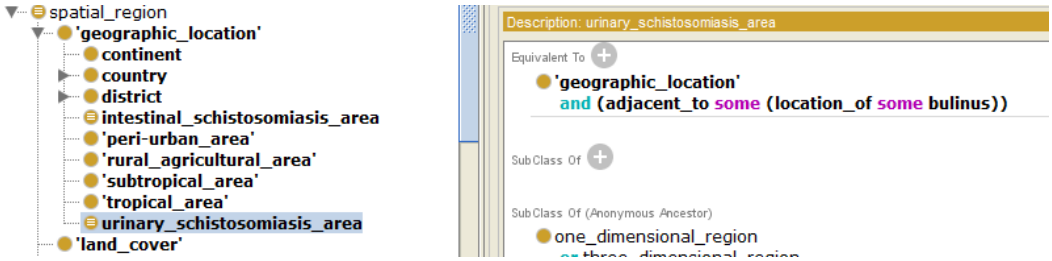


Figure 2 : Les zones à risque de schistosomiase urinaire

Tableau 2 : Métrique de l'ontologie IDOSCHISTO

Ontologie	#Concepts	#Relations	#Type	Total
Ontologie construite (concepts, relations et types propres ajoutés)				
IDOSCHISTO	58	14	0	72
Ontologies importées intégralement[6]				
Basic Formal Ontology (BFO)	39	0	0	39
Infectious Disease Ontology (IDO-Core)	285	14	0	299
ID Spreading Ontology (IDSDO-Core)	6	0	0	6
Relation Ontology (RO)	0	22	0	22
Ontologies importées partiellement				
Pathogen Transmission (TRANS)	4	0	0	4
Human Disease Ontology (DOID)	11	0	0	11
Population and Community Ontology (PCO)	21	28	0	49
NCBI Taxonomy Database	198	0	0	198
Ontology for Parasite LifeCycle (OPL)	130	6	0	136
Ontology of Medically Related Social Entities (OMRSE)	23	3	0	26
Environment Ontology (ENVO)	123	6	0	129
Exposure Ontology (EXO)	65	16	0	81
Total	958	109	0	1067

Les relations utilisées dans IDO-Core proviennent principalement de Relation Ontology (RO). Beaucoup d'autres relations sont importées à partir des ontologies réutilisées et nous en avons créé une quinzaine. Par exemple, parmi les relations ajoutées figurent « has_sign », « has_symptom », « has_vaccine », etc. liant les concepts que nous avons ajoutés. La relation « unfolds_in » décrivant le déroulement d'un processus dans l'espace a permis par exemple de dire que la propagation (« idsdo_spreading ») se propage dans (« unfolds_in) un espace (« geographical_location »). Une instanciation de cette relation serait qu'une épidémie (sous-

[6] Les concepts et relations obsolètes ne sont pas importés. Seuls les concepts de BFO et les relations de RO ne sont pas comptabilisés dans les chiffres indiquant les imports d'IDO-Core. Les autres concepts réutilisés par IDO-Core, par exemple à partir d'OBI, OGMS, etc., sont comptabilisés. IDSDO-Core ne comptabilise aussi que ses concepts et relations propres.

classe de propagation) se répand (unfolds_in) à Richard-Toll (une localité située au nord du Sénégal). Des classes définies ont été également proposées. Par exemple, les localités adjacentes à des points d'eau contenant le genre de mollusque bulinus sont des zones à risque de schistosomiase urinaire (cf. Figure 2). L'ontologie proposée est disponible à l'adresse suivante : https://github.com/gaoussoucamara/idoschisto/blob/master/idoschisto.owl. Les classes, relations et types de données importés sont présentés dans le Tableau 2.

Les modules correspondants aux différentes perspectives sont pour l'instant gérés par la création d'une annotation notée « *perspective* ». Elle prend trois valeurs possibles pour chaque concept et relation : *épidémiologique*, *biologique* ou *clinique*. Ainsi, chaque module peut être extrait en se fondant sur cette annotation. Cette manière de gérer des modules est encore triviale et son amélioration constitue l'une des perspectives de ce travail.

6. Cas d'étude : la schistosomiase à Richard Toll au Sénégal

Richard Toll est une zone située dans la vallée du fleuve Sénégal. Sa particularité dans notre étude est la densité de son réseau hydrographique (Traoré, 2000) qui est un facteur déterminant dans la transmission de la maladie. Les mollusques, hôtes intermédiaires des schistosomes, jouent un rôle important dans la transmission de la schistosomiase. Ainsi, les différentes espèces de mollusques rencontrées dans cette zone déterminent les types de schistosomiase rencontrés (Diaw et al., 1998). Nous recensons principalement deux grands genres dans la zone de Richard Toll : les *bulins* et les *biomphalaria.* Par conséquent, seuls deux parasites survivent dans la région, le *S. Haematobium* et le *S. Mansoni* causant respectivement la *schistosomiase urinaire* et la *schistosomiase intestinale.*

IDOSCHISTO est utilisée dans les études réalisées sur la schistosomiase à Richard Toll. Elle sert à l'annotation et au raisonnement sur les données. Les annotations portent sur les individus de la population, leurs répartitions géographiques, leurs activités, les points d'eau, les densités de mollusques et de parasites, les saisonnalités, les températures, les types de risque et les décisions associées, etc. Par exemple, la Figure 3 présente les annotations que nous avons introduites sur les données relatives aux quartiers de Richard-Toll et les définitions des relations d'*adjacence* de ces quartiers avec les différents points d'eau. Les relations d'adjacence sont particulièrement intéressantes car il est plus probable que les populations de ces quartiers privilégient ces points d'eau plutôt que ceux plus éloignés sauf si d'autres paramètres entrent en jeu (par exemple, aller pêcher dans les autres points d'eau car ce sont eux qui contiennent du poisson). Des annotations sur les espèces de mollusque (hôte intermédiaire des schistosomes) vivantes dans chaque point d'eau ont été introduites et les relations « has_intermediary_host » entre espèces de mollusque et espèces de schistosome ont été représentées.

Ces données sont issues des études conduites dans la localité et portent sur la répartition des espèces de mollusque, la localisation des points d'eau vis-à-vis des quartiers de la ville, etc. Par exemple, on peut exprimer la requête suivante en SPARQL : « *Quels sont les types de schistosomiases auxquelles sont exposées les populations du quartier de Ndiaw ?* » (cf. Figure 4). Son exécution donne les résultats dans la deuxième partie de la Figure 4. Il n'est pas déclaré dans la base de connaissances que les populations du quartier de Ndiaw sont exposées aux schistosomiases urinaire et intestinale mais c'est grâce aux connaissances représentées (concepts, relation, axiomes, instances) que ces résultats sont obtenus. En effet, on connaît quel type de schistosome cause quel type de schistosomiase, quel mollusque est hôte intermédiaire de quel schistosome, quel mollusque vit dans quel point d'eau et quel quartier est adjacent à quel point d'eau. L'analyse des URIs dans l'expression de la requête SPARQL montre la pertinence de la réutilisation des ontologies existantes et leurs utilisations effectives dans l'application visée.

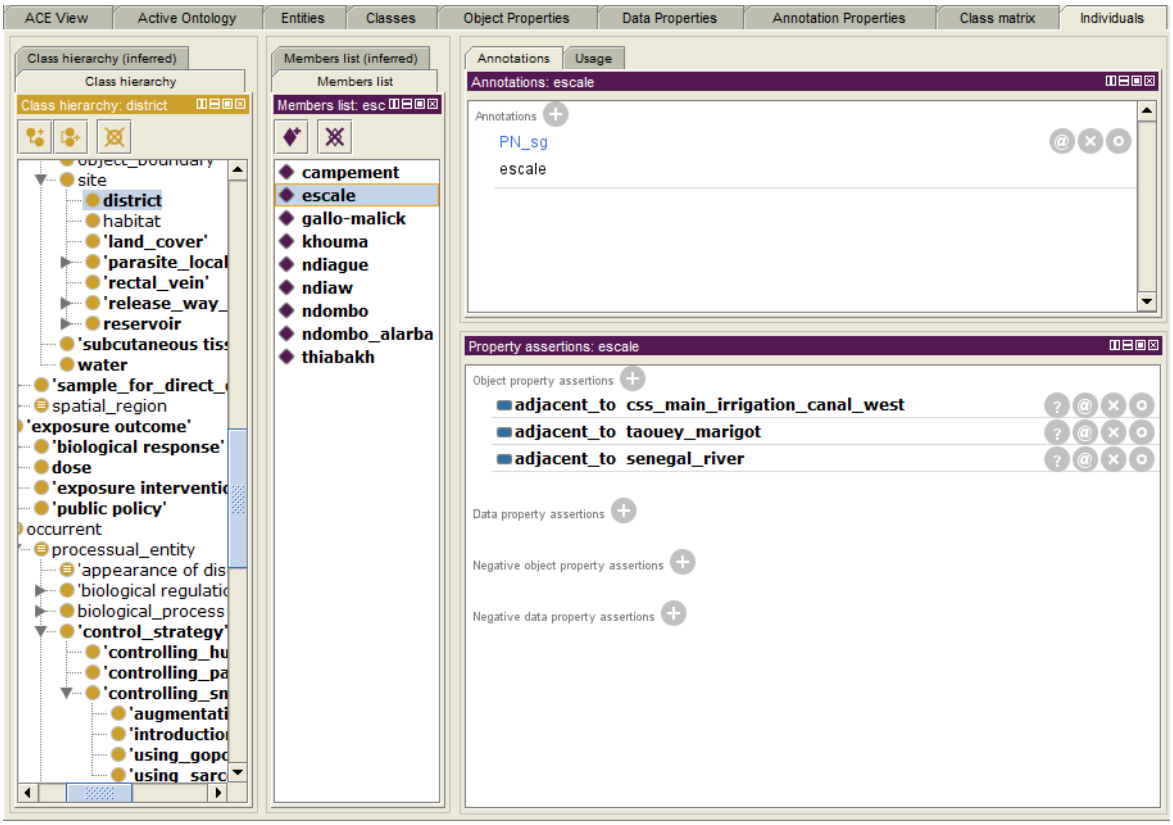


Figure 3 : Annotation des quartiers de Richard-Toll et définition de leurs relations d'adjacence avec les points d'eau

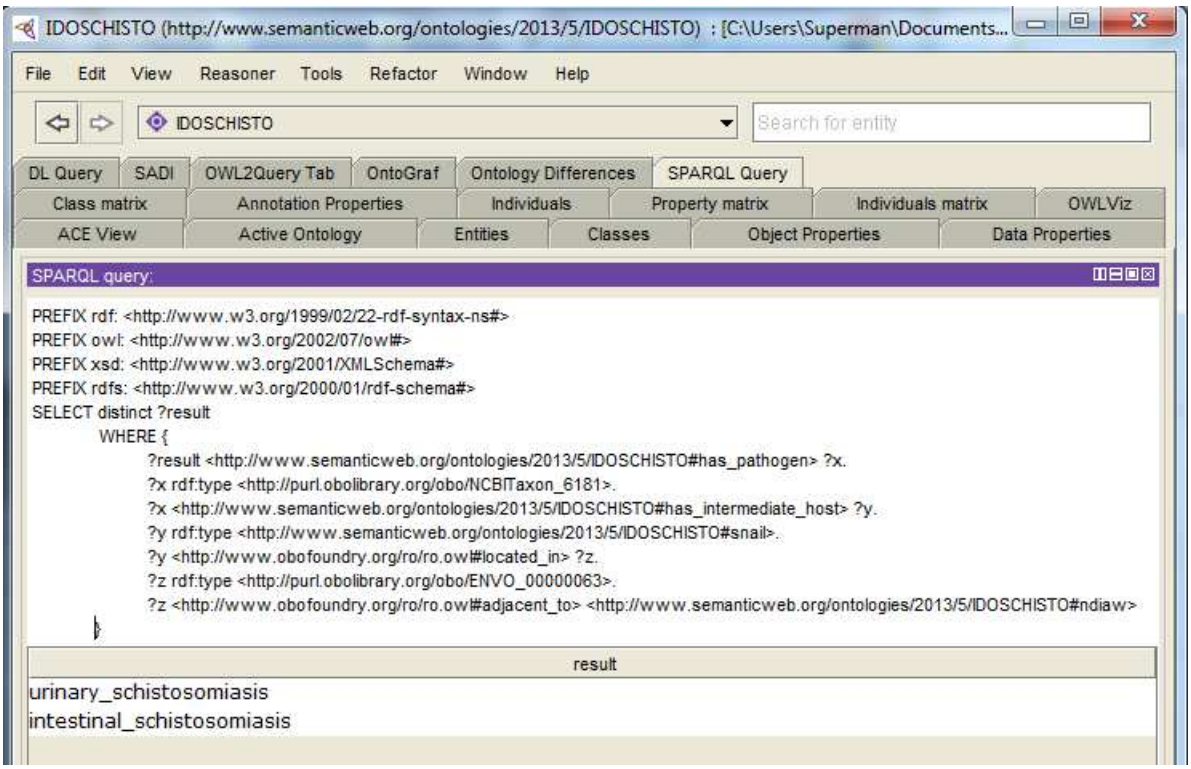


Figure 4 : Exemple de requête SPARQL dans IDOSCHISTO

7. Evaluation et validation de l'ontologie

La première phase de l'évaluation des ontologies de domaine a consisté à valider les modèles conceptuels avec les experts du domaine. Les modèles conceptuels de l'ontologie de la schistosomiase ont été examinés et validés avec des experts de la maladie au Sénégal. Cette validation ne garantit pas la couverture des connaissances du domaine mais elle valide au moins celles déjà représentées. Sur l'aspect formel, la réutilisation d'ontologie noyau ayant été déjà validée comme IDO-Core assure la cohérence d'IDOSCHISTO. Dans la réutilisation des ontologies de domaine spécifique, nous avons privilégié les ontologies qui ont connu un niveau de formalisation et d'implémentation avancées comme TRANS, OPL, PCO, NCBI, DOID, OGMS, etc. Un test a été effectué avec l'outil OOPS (Poveda-Villalón et al., 2012) pour détecter et corriger certaines incohérences et inconsistances. La validation par l'usage n'a pas pu se faire faute du retard dans l'implémentation du système de veille pour lequel l'ontologie est construite mais des requêtes SPARQL ont été testées avec l'interface de Protégé comme nous l'avons illustré dans la Figure 4.

8. Conclusion et discussion

Dans cet article, nous avons proposé une extension d'IDO-Core pour la construction de l'ontologie de domaine de la schistosomiase (IDOSCHISTO). La construction de cette ontologie suit l'approche que nous avons initiée dans (Camara et al., 2013). La conception de l'ontologie a été largement discutée avec les experts du domaine de la schistosomiase au Sénégal. L'ontologie du domaine de la schistosomiase étendant l'ontologie noyau du domaine des maladies infectieuses a été conçue pour couvrir l'ensemble des besoins en médecine. En effet, elle a été élaborée de façon modulaire de sorte qu'elle puisse être réutilisée partiellement pour des besoins spécifiques comme dans les systèmes d'information cliniques ou en recherche biologique. C'est pour cette raison que nous avons considéré dès le départ dans le cadre conceptuel la modularisation selon les perspectives épidémiologique, clinique et biologique. La construction d'IDOSCHISTO respecte également le principe de réutilisation de ressources ontologiques existantes dans OBO Foundry.

La modularité dans la partie domaine spécifique, notamment avec la distinction des différentes perspectives pose encore un réel problème. Bien que nous ayons créé et utilisé l'annotation *perspective* (avec *épidémiologique*, *clinique* et *biologique* comme valeurs possibles) sur un certain nombre de concepts clés, cette solution est loin d'être optimale. En effet, il y a également la partie concernant les axiomes qu'il faut gérer lors de l'extraction ou de la fusion des modules. L'évolution d'un module et son impact sur les autres modules et l'ontologie globale est également un défi à relever. En outre, la plupart des ontologies réutilisées sont encore au stade de *candidat* dans OBO Foundry. Elles sont par conséquent soumises à des évolutions plus ou moins importantes qu'il sera nécessaire de suivre et d'intégrer dans IDOSCHISTO.

Références

Camara, G., Despres, S., Djedidi, R., Lo, M., 2012a. Modélisation ontologique de processus dans le domaine de la veille épidémiologique, in: Actes de La Conférence RFIA 2012. Presented at the RFIA 2012 (Reconnaissance des Formes et Intelligence Artificielle).

Camara, G., Despres, S., Djedidi, R., Lo, M., 2014. Building a Schistosomiasis Process Ontology for an Epidemiological Monitoring System, in: Faucher, C., Jain, L.C. (Eds.), Innovations in Intelligent Machines-4, Studies in Computational Intelligence. Springer International Publishing, pp. 75–99.

Camara, G., Després, S., Djedidi, R., Lo, M., 2012b. Vers une ontologie des processus de propagation des maladies infectieuses, in: Actes des 23èmes journées francophones d'ingénierie des connaissances, IC 2012. Paris, France, pp. 99–114.

Camara, G., Després, S., Djedidi, R., Lo, M., 2013. Design of schistosomiasis ontology (IDOSCHISTO) extending the Infectious Disease Ontology. Presented at the In Proceedings of the 14th World Congress on Medical and Health Informatics, Copenhagen, Danmark.

Diaw, O.T., Vassiliades, G., Seye, M., Sarr, Y., 1998. Les mollusques hôtes intermédiaires des trématodoses humaines et animales : distribution et variation de densité dans les différents systèmes épidémiologiques de Richard-Toll, in: Hervé, J.-P., Brengues, J., Eau et Santé : Colloque, Dakar (SEN), 1994/11/14-15 (Eds.), Aménagements hydro-agricoles et santé (vallée du fleuve Sénégal), Colloques et Séminaires. ORSTOM, Paris, pp. 201–218.

Doran, P., Tamma, V., Iannone, L., 2007. Ontology module extraction for ontology reuse : an ontology engineering perspective. ACM Press, New York, pp. 61–70.

Falbo, R., Baião, F., Lopes, M., Guizzardi, G., 2010. The Role of Foundational Ontologies for Domain Ontology Engineering: An Industrial Case Study in the Domain of Oil and Gas Exploration and Production. Int J Inf Syst Model Des 1, 1–22.

Grenon, P., Smith, B., Goldberg, L., 2004. Biodynamic Ontology: Applying BFO in the Biomedical Domain, in: Stud. Health Technol. Inform. IOS Press, pp. 20–38.

Lin, Y., Xiang, Z., He, Y., 2011. Brucellosis Ontology (IDOBRU) as an extension of the Infectious Disease Ontology. J. Biomed. Semant. 2, 9.

Poveda-Villalón, M., Suárez-Figueroa, M.C., Gómez-Pérez, A., 2012. Validating Ontologies with OOPS!, in: Teije, A. ten, Völker, J., Handschuh, S., Stuckenschmidt, H., d' Acquin, M., Nikolov, A., Aussenac-Gilles, N., Hernandez, N. (Eds.), Knowledge Engineering and Knowledge Management, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 267–281.

Scherp, A., Saathoff, C., Franz, T., Staab, S., 2011. Designing core ontologies. Appl. Ontol. 6, 177–221.

Simons, P., Melia, J., 2000. Continuants and Occurrents. Proc. Aristot. Soc. Suppl. Vol. 74, 59–75+77–92.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S., 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. 25, 1251–1255.

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., Rosse, C., 2005. Relations in biomedical ontologies. Genome Biol. 6, R46.

Suárez-Figueroa, M.C., Gómez-Pérez, A., Fernández-López, M., 2012. The NeOn Methodology for Ontology Engineering, in: Suárez-Figueroa, M.C., Gómez-Pérez, A., Motta, E., Gangemi, A. (Eds.), Ontology Engineering in a Networked World. Springer Berlin Heidelberg, pp. 9–34.

Traoré, M., 2000. Importance des aménagements hydrauliques dans la transmission des schistosomoses, in: Chippaux, J.-P., Difficultés Rencontrées dans la Mise en Oeuvre des Programmes de Lutte contre les Schistosomoses en Afrique de l'Ouest : Atelier, Niamey (NER), 2000/02/15-18 (Eds.), La lutte contre les schistosomoses en Afrique de l'Ouest, Colloques et Séminaires. IRD, Paris, pp. 23–29.

Xiang, Z., Courtot, M., Brinkman, R.R., Ruttenberg, A., He, Y., 2010. OntoFox: web-based support for ontology reuse. BMC Res. Notes 3, 175.

Validation de la sémantique d'un langage iconique médical à l'aide d'une ontologie : méthodes et applications⋆

Jean-Baptiste Lamy[1], Lina F. Soualmia[1,2], Alain Venot[1], Catherine Duclos[1]

[1] LIMICS, Université Paris 13, Sorbonne Paris Cité, 93017 Bobigny, France, INSERM UMRS 1142, UPMC Université Paris 6, Sorbonne Universités, Paris
{jean-baptiste.lamy,alain.venot}@univ-paris13.fr, catherine.duclos@avc.ap-hop-paris.fr

[2] TIBS, LITIS EA 4108, Institute of Biomedical Research, Rouen, France
lina.soualmia@chu-rouen.fr

Résumé : Les langages iconiques permettent de représenter des concepts par la combinaison de primitives graphiques (couleurs, pictogrammes...). Les exemples sont nombreux, des panneaux routiers aux icônes des interfaces informatiques. Cependant, ces langages ne définissent pas la sémantique de leurs icônes, ce qui pose plusieurs problèmes : combinaisons inconsistantes de pictogrammes, interprétations différentes d'une même icône par deux personnes différentes, difficultés à aligner les icônes avec des ressources existantes...

Le langage iconique VCM (Visualisation des Concepts en Médecine) permet de représenter par des icônes les principaux antécédents, maladies, traitements,... Dans cet article, nous proposons de valider la sémantique du langage iconique VCM à l'aide d'une ontologie. Trois applications de cette ontologie sont décrites : la vérification de la consistance des icônes constituées, l'alignement semi-automatique avec une terminologie médicale, et la génération d'un lexique des pictogrammes.

Mots-clés : Icônes, Langage iconique, Ontologies, Alignement, Médecine

1 Introduction

Il est bien connu qu'un "bon schéma vaut mieux qu'un long discours". C'est pourquoi de nombreux icônes, symboles ou pictogrammes ont été proposés (Dreyfuss, 1984) dans des domaines aussi variés que les interfaces homme-machine, la signalisation dans les lieux publics ou l'étiquetage des produits chimiques. Cependant le nombre d'icônes qu'un être humain peut mémoriser n'est pas infini et, lorsque l'on veut représenter un grand nombre de concepts, il n'est pas possible d'avoir une icône par concept. Une solution est alors de concevoir un *langage iconique*, dans lequel un grand nombre d'icônes peut être créé en combinant un nombre restreint de primitives tels que des couleurs ou des pictogrammes, en suivant une grammaire spécifique. Un exemple bien connu est la signalisation routière, où les panneaux sont composés de plusieurs éléments (cercle rouge, pictogramme,...). VCM (Visualisation des Concepts en Médecine, Lamy *et al.* 2008, 2009) est un langage iconique permettant de représenter les principaux concepts médicaux (maladies, traitements,...) par des icônes. Ce langage a pour objectif de faciliter l'accès des professionnels de santé aux documents médicaux. En effet, le volume de données et de connaissances textuelles

⋆. Ce travail a été financé par l'ANR (Agence Nationale de la Recherche) au travers des projets de recherche L3IM (ANR-08-TECS-007) et SiFaDo (ANR-11-TECS-0014).

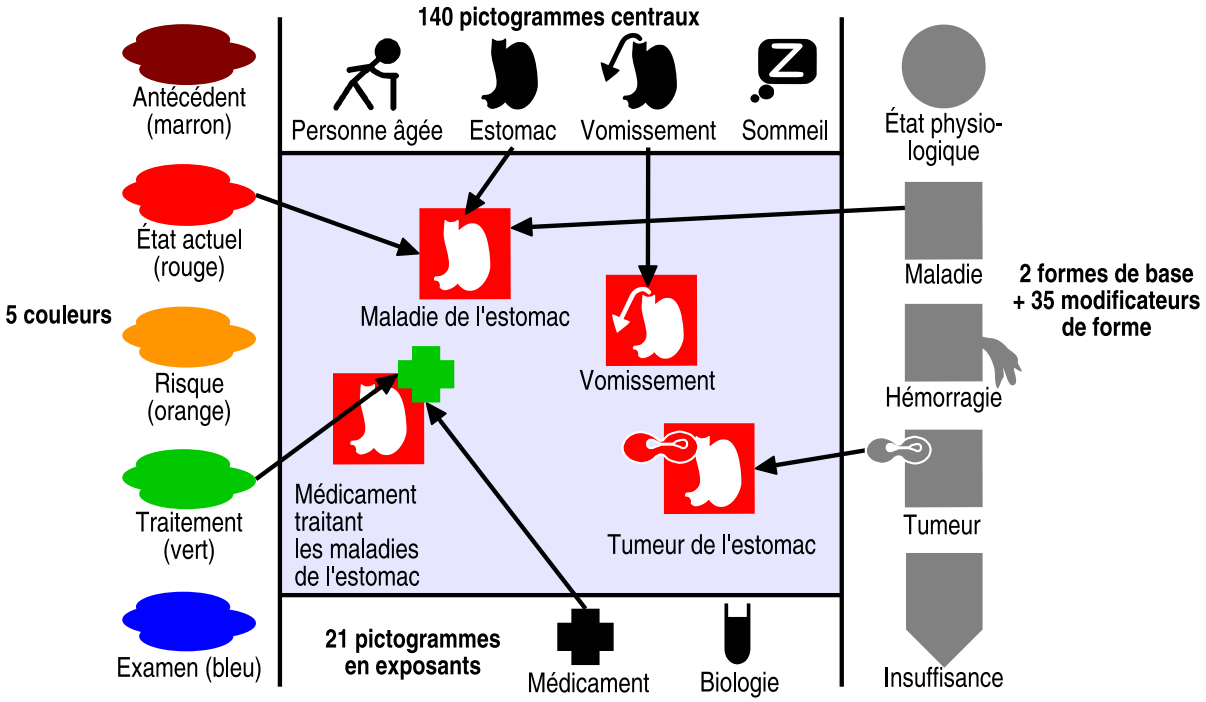


FIGURE 1 – Exemples d'icônes VCM générées par la combinaison de plusieurs éléments.

rend difficile la lecture des dossiers patients, des Résumés des Caractéristiques Produits (RCP) des médicaments ou des guides de bonnes pratiques cliniques (Ely *et al.*, 2002).

Cependant, dans un langage iconique, une icône syntaxiquement correcte peut malgré tout être sémantiquement fausse : par exemple un panneau routier avec le triangle rouge *attention danger* et le pictogramme *pneu neige*. L'absence de sémantique validée rend aussi plus difficile l'alignement des icônes avec les ressources ou terminologies existantes du domaine : les alignements doivent être réalisés manuellement par un expert qui interprète les icônes, mais comme toute interprétation humaine, celle-ci risque d'être subjective.

Dans cet article, nous proposons une approche consistant à valider la sémantique d'un langage iconique à l'aide d'une ontologie formelle. Nous présenterons le langage VCM qui servira d'exemple, puis la construction de l'ontologie des icônes VCM. Ensuite, nous décrirons trois applications de cette ontologie : la vérification de la consistance des icônes constituées, l'alignement semi-automatique avec une terminologie médicale de référence, et la génération d'un lexique des pictogrammes. Enfin, nous discuterons des avantages et des limites de cette approche fondée sur une ontologie formelle. Les ontologies, alignements et programmes réalisés au cours de ce travail ont été diffusés en logiciel libre dans PyMedTermino (https://pypi.python.org/pypi/PyMedTermino), un ensemble de modules pour accéder aux terminologies médicales en langage Python.

2 Le langage VCM

Le langage iconique VCM (Lamy *et al.*, 2008, 2009) propose des icônes pour représenter les principales conditions cliniques d'un patient, dont les symptômes, les maladies, les états physiologiques (tels que la grossesse ou les classes d'âge), les risques et les antécédents de maladie, les traitements médicamenteux ou non, les examens de biologie et les procédures de suivi. VCM comprend un ensemble de primitives graphiques (pictogrammes, formes et couleurs) pouvant être combinées selon une grammaire pour générer un grand nombre d'icônes. Ces icônes n'ont pas pour objectif d'être aussi précises que les textes médicaux, mais visent au contraire à les compléter de manière synthétique. Un didacticiel est disponible sur le site Internet dédié à VCM : http://vcm.univ-paris13.fr/.

La Figure 1 montre différents exemples de combinaisons. Une icône VCM comprend une couleur, une forme de base et un ensemble de modificateurs de forme, un pictogramme central, ainsi qu'éventuellement un exposant associant un pictogramme et une couleur, et un second exposant. Une icône simple combine (1) une couleur indiquant l'aspect temporel de l'icône : rouge pour un état actuel du patient, marron pour un antécédent, orange pour un risque futur ; (2) une forme de base : un cercle pour un état physiologique ou un carré pour un état pathologique ; et (3) un pictogramme central indiquant la localisation anatomo-fonctionnelle (par exemple *cardiaque*, *pulmonaire*,...) ou la caractéristique patient (par exemple *grossesse*) impliquée. Les structures anatomiques et les fonctions qu'elles réalisent sont représentées par le même pictogramme (par exemple *poumon* et *respiration*).

Les icônes de maladie ou de symptôme peuvent ensuite être précisées, selon deux approches différentes : (1) pour les maladies ou symptômes spécifiques à un système anatomo-fonctionnel (par exemple les *vomissements*, qui sont spécifiques à l'*estomac*), le pictogramme central est modifié ; (2) pour les maladies ou symptômes génériques pouvant être décrits comme une morphologie associée à un système anatomo-fonctionnel (par exemple les *tumeurs*, les *infections* ou les *insuffisances* d'une fonction), un modificateur de forme est ajouté à la forme de base (par exemple, les tumeurs sont représentées par deux cellules en division). Ces deux approches peuvent être combinées, et plusieurs modificateurs de forme peuvent être présents du moment qu'ils ne se recouvrent pas spatialement.

Les icônes de traitements et d'examens sont construites à partir des icônes de la maladie traitée ou du risque surveillé, en ajoutant un pictogramme en exposant, de couleur verte pour les traitements ou bleue pour les examens. Ce pictogramme en exposant indique le type de traitement (par exemple *chirurgical*) ou d'examen (par exemple *imagerie*). Un second exposant peut être ajouté pour représenter un professionnel de santé ou un document en relation avec une maladie, par exemple l'icône *cardiologue* sera construite en ajoutant l'exposant *professionnel de santé* à l'icône des *maladies cardiaques*.

3 L'ontologie des icônes VCM

L'ontologie des icônes VCM (Lamy *et al.*, 2013a) a pour objectif d'aider à la validation de la sémantique du langage VCM. Nous présenterons tout d'abord les principes généraux suivis lors de la construction de cette ontologie, puis nous la décrirons plus en détails.

3.1 Principes généraux pour la construction de l'ontologie

Le premier principe que nous avons appliqué lors de la construction de l'ontologie a été de distinguer les primitives de VCM et leur signification : en effet le *pictogramme en forme de poumon* est distinct de l'organe *poumon*. Un second principe a été d'utiliser au maximum les relations de subsomption *est-un* plutôt que d'autres relations comme les relations méréologiques *partie-de*, puisque la plupart des outils d'édition et des moteurs d'inférences s'appuient sur la subsomption. En particulier, les objets anatomiques ont été désignés comme "structure + adjectif" plutôt que par leur nom d'organe ; nous dirons par exemple "une *structure bronchique* est une *structure pulmonaire*" plutôt que "les *bronches* sont une partie des *poumons*". Les organes peuvent ensuite être ajoutés comme fils des structures anatomiques : le *poumon* est une *structure pulmonaire*. Bien que contraire au principe du "biais minimal d'encodage", on retrouve cette approche dans plusieurs terminologies médicales, dont la SNOMED CT (*Systematized Nomenclature of Medicine - Clinical Terms* Cornet 2009) et nous l'avons donc adoptée. Un troisième principe est que toute icône décrit un état du patient, y compris pour les icônes de traitement ou d'examen. Par exemple l'icône *anti-asthmatique* a été modélisée comme *patient traité par un anti-asthmatique*. Cela correspond à la manière dont VCM représente les traitements et les examens, en reprenant la pathologie qui est traitée ou le risque qui est surveillé.

L'ontologie intègre deux types de contraintes : *(i) des contraintes graphiques* portant sur les primitives de VCM : par exemple les modificateurs de forme *tumeur* et *virus* occupent le même emplacement sur les icônes, et ne peuvent donc pas être conjointement associés ; et *(ii) des contraintes médicales* portant sur les structures anatomiques, les morphologies,... représentées par les primitives : par exemple une *tumeur* est une morphologie qui ne peut s'appliquer qu'à une structure anatomique, mais pas à une fonction biologique.

Lors de la construction de l'ontologie, un jeu d'une centaine d'icônes tests a été mis au point de manière itérative, et un raisonneur a été utilisé pour tester la consistance de l'ontologie et vérifier les résultats obtenus sur les icônes tests.

3.2 Structure de l'ontologie des icônes VCM

L'ontologie des icônes VCM (Figure 2, voir Lamy *et al.* 2013a pour une version plus détaillée) a été découpée en trois modules. Le premier (240 concepts, 21 relations et 2 597 axiomes) décrit les primitives et les icônes VCM. Il a été généré à partir de la liste des primitives du langage VCM. Le fichier OWL obtenu a ensuite été importé puis édité manuellement dans Protégé pour y ajouter les règles de composition des icônes et les contraintes graphiques. Ces règles restreignent les composantes d'une icône (par exemple, au plus un pictogramme central) et prennent en compte les contraintes spatiales.

Le second module (369 concepts, 18 relations et 828 axiomes) décrit les concepts médicaux représentés par les primitives du langage VCM : structures anatomiques, fonctions biologiques, morphologies processus pathologiques, caractéristiques patient (telles que les classes d'âge) et types de traitements et de surveillances. L'ontologie contient les concepts médicaux de base et les règles de combinaison, mais pas l'ensemble des maladies, traitements,... qui peuvent être générées par combinaison (il s'agit donc de *post-coordination*). Ce second module a été modélisé avec Protégé, en s'appuyant sur les terminologies médicales (SNOMED CT en particulier) et le réseau sémantique de l'UMLS (Kashyap &

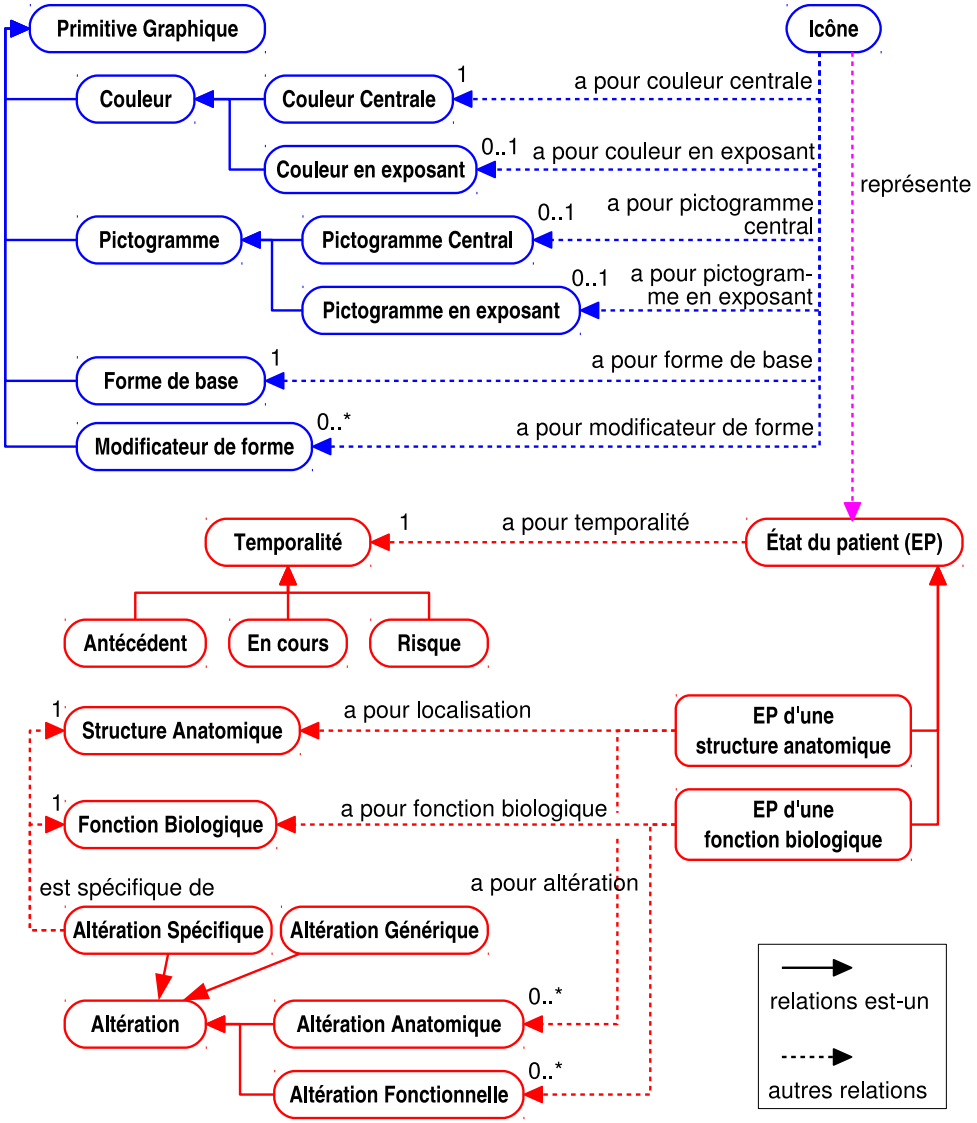


FIGURE 2 – Principaux concepts et relations dans l'ontologie des icônes VCM. Les couleurs correspondent aux modules de l'ontologie.

Borgida, 2003), en se limitant au niveau de granularité élevé qui est celui de VCM.

Les concepts de ces deux modules sont reliés par des relations *représente* (509 axiomes). Par exemple le pictogramme central *poumon* est seulement présent sur des icônes qui représentent des concepts médicaux liés à une *structure pulmonaire* ou à la *fonction respiratoire*. Ces relations ont été générées automatiquement à partir d'un fichier texte faisant correspondre à chaque primitive le ou les concepts médicaux correspondants.

L'ontologie comprend en tout 609 concepts, 41 relations et 3 934 axiomes, définis à l'aide du langage OWL-DL. Elle appartient à la famille *ALCRIQ* des logiques de description (Attribute Language, Complex concept negation, Role, Inverse property, Qualified cardinality restriction), qui est décidable (Horrocks & Sattler, 2003).

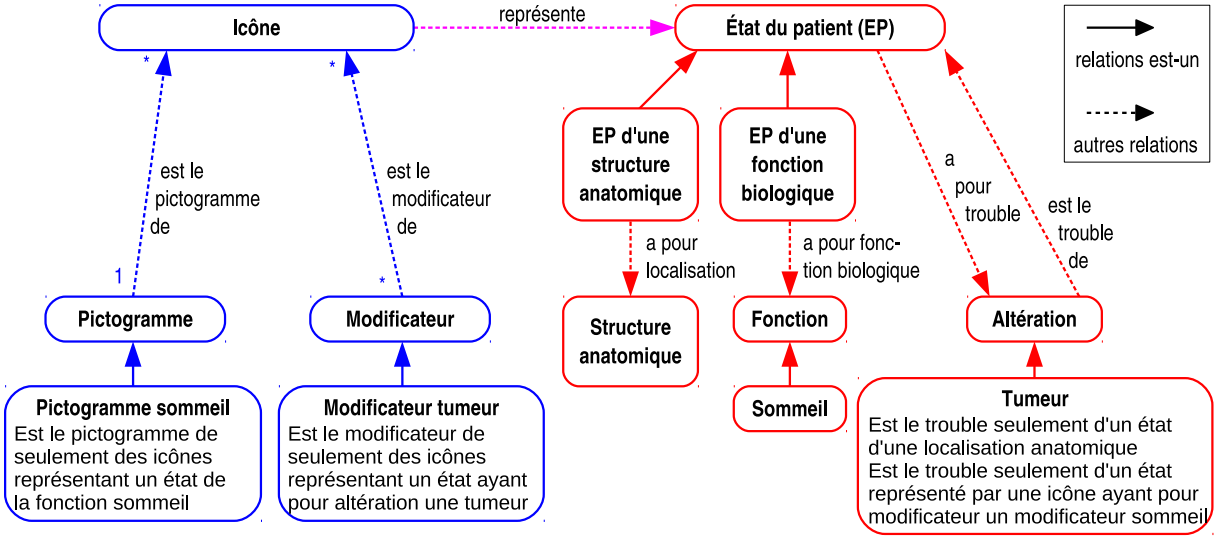


FIGURE 3 – Représentation de l'icône inconsistante *tumeur du sommeil* dans l'ontologie.

4 Vérification de la consistance des icônes

La première application de l'ontologie des icônes VCM a été la vérification de la consistance des icônes (Lamy *et al.*, 2013a). En effet, certaines combinaisons du langage VCM sont inconsistantes : par exemple une icône associant le modificateur de forme *tumeur* et le pictogramme central *sommeil* signifie "tumeur du sommeil", ce qui n'a aucun sens du point de vue médical. Cependant, ces icônes inconsistantes peuvent poser problème, notamment lorsque des utilisateurs sont amenés à construire eux-mêmes une icône en sélectionnant plusieurs primitives. Nous allons donc voir comment les contraintes exprimées dans l'ontologie permettent de vérifier la consistance des icônes.

La Figure 3 montre un exemple de représentation dans l'ontologie d'une icône inconsistante. L'inconsistance peut être déduite à partir des contraintes modélisées dans l'ontologie : (a) l'icône a le pictogramme *sommeil*, et donc elle représente un état de la fonction *sommeil*, (b) l'icône a le modificateur de forme *tumeur*, et donc elle représente un état qui a pour altération la morphologie *tumeur*, (c) la morphologie *tumeur* est une altération anatomique, qui ne peut s'appliquer qu'à des états d'une structure anatomique, (d) les structures anatomiques et les fonctions biologiques sont disjointes.

Ce raisonnement a pu être reproduit avec le raisonneur HermiT (Motik *et al.*, 2009), soit pour une seule icône, soit pour une classe d'icônes, en créant dans l'ontologie le concept des icônes partageant telle ou telle composante. La détermination des inconsistances a ensuite été évaluée (Lamy *et al.*, 2013a).

5 Alignement de VCM avec les terminologies médicales : la SNOMED CT

L'utilisation de VCM dans les applications médicales nécessite son alignement avec des référentiels existants, afin notamment de pouvoir associer automatiquement des icônes aux ressources déjà indexées par ces référentiels, comme par exemple les Dossiers Pa-

tient Informatisés. Dans cette section, nous étudierons la SNOMED CT, une terminologie couvrant les différents concepts médicaux, dont l'anatomie, les conditions cliniques, les procédures,... Elle inclut de nombreuses relations entre ces concepts : des relations de subsomption *est-un* mais aussi des relations entre conditions cliniques et structures anatomiques ou morphologies associées (par exemple l'hépatite *est localisée dans* le foie et *a pour morphologie* l'inflammation).

La méthode que nous proposons pour l'alignement s'appuie sur la nature compositionnelle similaire de la SNOMED CT et de l'ontologie des icônes VCM. L'alignement s'est fait en deux étapes : (1) alignement manuel des concepts médicaux de l'ontologie des icônes VCM (n=369) aux termes SNOMED CT correspondants (structures anatomiques, étiologies, etc), et (2) alignement automatique des termes SNOMED CT de conditions cliniques aux icônes VCM, en décomposant les termes SNOMED CT et en traduisant chacun des termes obtenus avec l'alignement manuel précédent. Cette méthode a initialement été appliquée à un sous-ensemble de la SNOMED CT, la CORE problem list (Lamy *et al.*, 2013b) ; nous présenterons ici les résultats sur l'ensemble de la SNOMED CT.

5.1 Alignement manuel entre concepts médicaux de l'ontologie VCM et la SNOMED CT

Les 369 concepts de l'ontologie des icônes VCM ont été manuellement alignés sur la SNOMED CT. La SNOMED CT utilisant l'héritage multiple, certaines structures anatomiques sont classées dans plusieurs branches. Cela conduit à associer plusieurs pictogrammes à ces structures, et donc plusieurs icônes aux maladies correspondantes. Par exemple les *osselets de l'oreille* sont à la fois une *structure osseuse* et une *structure auditive*, et donc une *maladie des osselets* aura les deux icônes *maladie de l'os* et *maladie de l'oreille*. Cependant, si cela est juste d'un point de vue ontologique, cela ne correspond pas à ce qu'attend un clinicien : en effet, les maladies sont classées par spécialité notamment durant les études médicales, chaque maladie étant attribuée de manière consensuelle à une et une seule spécialité (par exemple les maladies des osselets relèvent de l'ORL, oto-rhino-laryngologie, ce qui fait qu'un médecin voyant l'icône *maladie de l'os* ne pensera pas à une maladie des osselets). C'est pourquoi nous avons choisi d'associer chaque structure anatomique à au plus un pictogramme (l'oreille dans notre exemple).

Pour cela, une première version de l'alignement manuel a été réalisée, puis nous avons recherché tous les termes SNOMED CT de structures anatomiques qui conduisaient *via* l'ontologie à plusieurs pictogrammes (n=181). Chacun de ces termes a ensuite été associé à un et un seul concept de l'ontologie, ce qui a conduit à la création de nouveaux concepts médicaux dans l'ontologie (n=97, inférieur à 181 car certains concepts proches ont été regroupés). Ces concepts ont ensuite été associés à un seul pictogramme, choisi selon (1) la spécialité médicale associée à la structure anatomique et aux maladies correspondantes, et (2) la position de ces maladies dans les terminologies médicales monoaxiales et notamment la CIM10 (Classification Internationale des Maladies, version 10).

Au total, l'alignement manuel a fait intervenir 1 753 termes SNOMED CT.

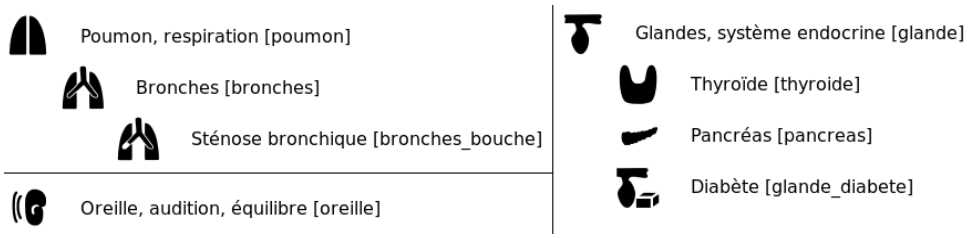


FIGURE 4 – Trois extraits du lexique des primitives VCM rédigé manuellement. Les identifiants des primitives figurent entre crochets.

5.2 Alignement automatique entre SNOMED CT et VCM

Les termes de conditions cliniques de la SNOMED CT ont ensuite été alignés automatiquement aux icônes VCM, en procédant par décomposition. Chaque terme est décomposé à l'aide des relations de la SNOMED CT, puis les termes SNOMED CT obtenus sont traduits en concept de l'ontologie VCM à l'aide de l'alignement précédent, et en prenant en compte les relations *est-un* et *partie-de* existant dans la SNOMED CT. Ces concepts sont ensuite traduits en primitives VCM à l'aide des relations de l'ontologie VCM, puis les primitives obtenues sont assemblées pour former une icône. Par exemple, le terme SNOMED CT *Uveitis* est décomposé en une structure anatomique, *Uveal tract*, et une morphologie, *Inflammation*. *Uveal tract* est une partie de *Entire eye* qui est une *Structure of visual system*, laquelle est traduite en *Structure visuelle* dans l'ontologie VCM. *Structure visuelle* est alors traduit en *pictogramme œil* dans les primitives VCM. *Inflammation* conduit au modificateur de forme *carré avec une flamme*, et ce modificateur est combiné au pictogramme *œil* pour donner l'icône adéquate.

L'alignement obtenu porte sur l'ensemble des 99 626 conditions cliniques (*clinical findings*) de la SNOMED CT, et a fait appel à 1 957 icônes VCM. 77 754 (78,0%) ont été alignés avec une seule icône, 7 573 (7,6%) ont été alignés avec 2 icônes et 517 (0,5%) avec 3 icônes ou plus ; 13 782 (13,8%) termes n'ont pas pu être alignés avec VCM (la méthode aboutissant à aucune icône ou à une icône "vide" sans pictogramme ni modificateur de forme). L'analyse manuelle des termes non-alignés montre qu'il s'agit principalement de termes qui ne sont pas des conditions cliniques au sens de VCM (par exemple *Drug therapy finding*), de termes servant à qualifier des conditions cliniques (*Clinical stage finding*), de termes très généraux (*Alive*) ou de symptômes non-associés à une localisation précise (*Erythema*, à distinguer de *Erythema of skin*).

6 Génération du lexique des primitives du langage VCM

La documentation du langage VCM comprend un lexique des primitives du langage (pictogrammes, modificateurs de forme et couleurs). Ce lexique sert à l'apprentissage de VCM, mais aussi de référence pour les experts. Le lexique se présente sous la forme d'une liste arborescente, associant sur chaque ligne une primitive et le ou les libellés associés, ainsi que l'identifiant de la primitive.

La version originelle du lexique (figure 4) a été rédigée manuellement. Cette rédaction manuelle pose plusieurs problèmes : le lexique doit être remis à jour systématiquement à

chaque modification du langage iconique, et il peut subsister des ambiguïtés (par exemple deux pictogrammes pouvant correspondre au même organe) et des zones d'ombre (par exemple un organe pour lequel aucun pictogramme n'est indiqué).

Dans cette section, nous proposons une méthode pour générer automatiquement le lexique à partir de l'ontologie des icônes VCM.

6.1 Méthode de génération du lexique

La construction du lexique s'est faite en quatre étapes : (1) l'extraction de l'ensemble des primitives VCM de l'ontologie , (2) pour chaque primitive, récupération *via* les relations de l'ontologie de la liste des concepts médicaux que la primitive peut représenter, (3) ordonnancement de ces concepts et (4) obtention du ou des libellés associés à chaque concept, et génération du lexique.

Lors de l'étape 2, une primitive correspond souvent à plusieurs concepts. En effet, VCM utilise fréquemment le même pictogramme pour représenter un organe et sa fonction. La liste de concepts inclut aussi tous les concepts descendants (fils, petit-fils,... par exemple *structure de la plèvre* est un descendant de *structure pulmonaire*), à l'exception de ceux qui sont reliés à une primitive plus spécifique (par exemple *structure bronchique* qui est relié au pictogramme *bronche*).

Lors de l'étape 3, lorsque plusieurs concepts sont associés à une même primitive, se pose alors la question de l'ordre dans lequel ils apparaîtront dans le lexique. Pour cela, nous avons mis au point des règles combinant la nature des concepts (anatomique ou fonctionnelle), le niveau d'échelle (macroscopique ou microscopique), la spécificité (générale ou spécifique) et la lisibilité des libellées correspondants (les noms d'organe sont généralement plus lisibles que les noms de structure correspondants, par exemple *poumon* vs. *structure pulmonaire*). Les règles appliquées pour ordonner les concepts sont les suivantes (par ordre décroissant de priorité) :

1. Lorsqu'un organe est présent parmi les concepts, celui-ci vient en premier.
2. Les concepts appartenant à plusieurs hiérarchies anatomiques (par exemple les *osselets de l'oreille interne* qui sont à la fois dans la hiérarchie des *structures auditives* et dans la hiérarchie des *structures osseuses*) sont placés en dernier, et écrits en gris.
3. Les concepts de structures anatomiques sont placés avant les concepts de fonctions biologiques.
4. Les concepts de structures anatomiques sont ordonnés entre eux en allant du macroscopique au microscopique. Pour cela, nous avons utilisé les quatre niveaux d'échelle (région anatomique, tissus, cellule, liquide) que distingue l'ontologie.
5. Les concepts généraux sont placés avant les concepts plus spécifiques (en s'appuyant sur les relations *est-un*, par exemple *diabète* est placé avant *diabète de type 2*).

Lors de l'étape 4, nous avons récupéré les libellés associés aux concepts. Pour chaque concept, l'ontologie contient un libellé principal, et éventuellement un ou plusieurs synonymes ou hyponymes. Le libellé principal a été placé en premier, suivi de la mention "inclut" et des autres libellés. L'ordre dans lequel les primitives apparaissent dans le lexique a été déterminé à la main, en reprenant l'ordre du lexique manuel, qui suivait une logique anatomique (par exemple en regroupant le système cardiovasculaire, pulmonaire,...).

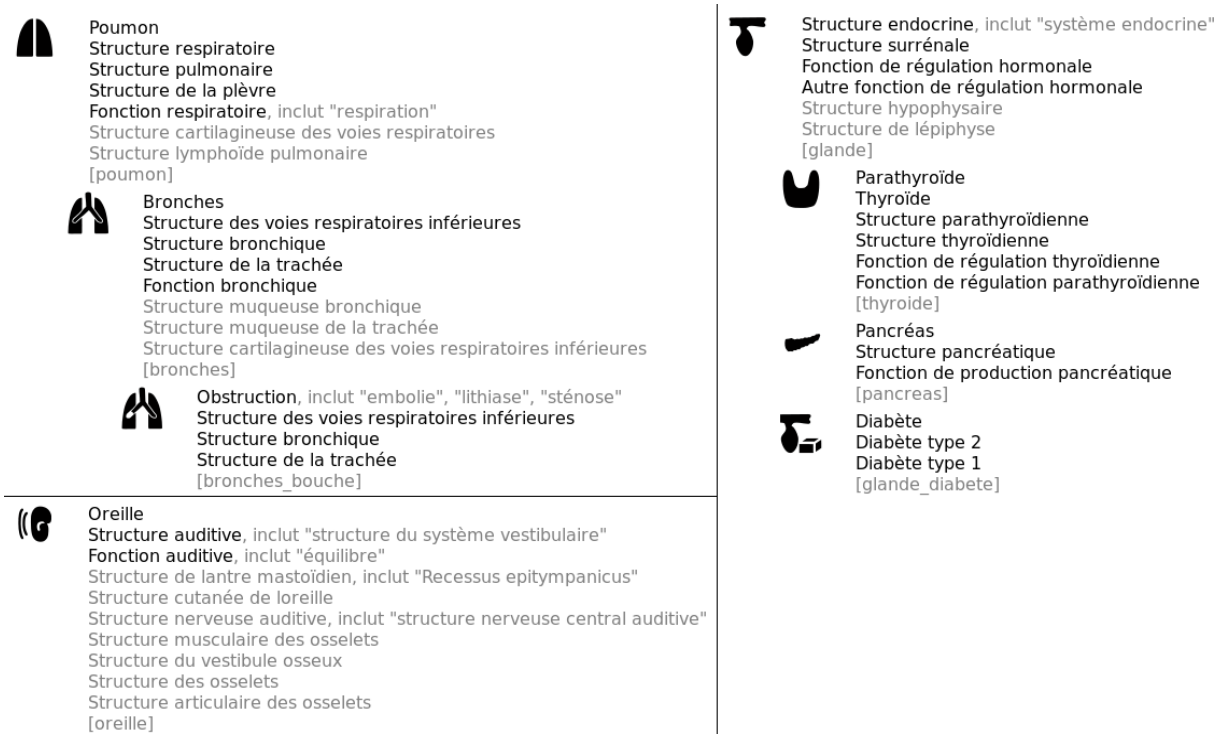


FIGURE 5 – Trois extraits du lexique des primitives VCM produit à partir de l'ontologie.

6.2 Le lexique produit à partir de l'ontologie

La Figure 5 montre des extraits du lexique produit à partir de l'ontologie. En comparaison avec l'ancien lexique (Figure 4), le nouveau lexique est plus riche.

L'alignement des concepts médicaux de l'ontologie des icônes VCM avec la SNOMED CT garantit la couverture de l'ensemble du domaine, notamment en ce qui concerne l'anatomie. Le lexique généré ne contient donc pas de zone d'ombre : toutes les structures anatomiques sont présentes dans le lexique, soit directement, soit via une structure parente plus générale (par exemple *structure de la plèvre* a été rattachée au pictogramme poumon et apparaît sur le lexique, alors qu'elle était absente de l'ancien lexique). De plus, lors de l'alignement avec la SNOMED CT, nous avons repéré les 181 concepts médicaux ambigus qui étaient associés à plusieurs primitives. Tous ces concepts ambigus ont été ajoutés dans l'ontologie et apparaissent donc dans le lexique, clairement associés à un seul pictogramme (par exemple *structure des osselets*, qui apparaît en face du pictogramme *oreille*).

7 Discussion et conclusion

Nous avons vu comment valider la sémantique d'un langage iconique en représentant les concepts relatifs aux icônes à l'aide d'une ontologie formelle, puis nous avons proposé trois applications d'une telle ontologie : la vérification de la consistance des icônes, l'alignement des icônes avec des terminologies existantes et la génération d'un lexique des pictogrammes. Les langages iconiques sont traditionnellement définis par la décomposition

graphique des composantes des icônes (Meunier, 1998), plus récemment X. Ma a proposé de structurer selon le modèle Hyppertopic, issu du Web socio-sémantique, des icônes destinées au taguage de documents (Ma & Cahier, 2012). La validation ontologique que nous proposons ici va plus loin en séparant les éléments graphiques (le signifiant) de ce qu'il représente (le signifié), et a permis d'automatiser les différentes applications. Les trois applications présentées auraient pu être accomplies avec des représentations des connaissances spécifiques à chacune d'elles (par exemple des règles de grammaire pour vérifier la syntaxe des icônes), cependant l'ontologie a permis de toutes les réaliser.

Ces méthodes ont été appliquées avec succès à VCM, un langage iconique représentant les principaux concepts médicaux. Vu la généricité de l'approche proposée, elles pourraient cependant être appliquées à d'autres langages iconiques, tels que les panneaux routiers.

La première application de l'ontologie a porté sur la vérification de la consistance des icônes (Lamy *et al.*, 2013a). Ensuite, nous avons aligné les icônes VCM avec une terminologie de référence en médecine, la SNOMED CT (Lamy *et al.*, 2013b). Cet alignement a pu être réalisé de manière semi-automatique grâce aux relations présentes dans l'ontologie et dans la SNOMED CT, cependant la méthode utilisée sera difficilement reproductible avec des ressources terminologiques moins structurées ou formalisées, telle que la CIM10. Enfin, nous avons produit un lexique des primitives graphiques à partir de l'ontologie d'un langage iconique. Le lexique ainsi construit est plus riche qu'un lexique rédigé à la main, et l'alignement de l'ontologie avec les terminologies du domaine médical permet de s'assurer de l'absence de zones d'ombre ou d'ambiguïté dans le lexique. Ce lexique peut être mis à jour automatiquement en cas de modification de l'ontologie.

Les principales difficultés rencontrées lors de la génération du lexique étaient liées à l'ordre des éléments présents dans le lexique. En effet, l'ontologie ne définit pas d'ordre entre les concepts partageant une relation (par exemple si l'ontologie définit la *bouche*, l'*œsophage* et l'*estomac* comme étant des *structures digestives*, elle ne donne pas l'ordre dans lequel présenter ces trois organes). Nous avons proposé des règles pour ordonner les différents concepts listés en face d'une entrée du lexique, en prenant en compte la nature, le niveau d'échelle, la spécificité, et la lisibilité des libellés des concepts. En revanche, l'ordre dans lequel figure les primitives dans le lexique a été déterminé manuellement. Cet ordre est important car un utilisateur s'attend à trouver ensemble les primitives d'un même système (système digestif par exemple). De plus, pour certains systèmes comme le système digestif, il existe un ordre logique dans lequel présenter les primitives, en suivant le trajet du bol alimentaire (bouche, œsophage, estomac, intestin grêle, côlon, anus). Des ontologies plus complètes sur l'anatomie, telles que la FMA (Rosse & Mejino JL, 2003), contiennent des relations de connexion entre organes (par exemple la bouche est connectée à l'œsophage), à partir desquelles nous pourrions inférer l'ordre des organes du tube digestif. Cependant, même ainsi, il n'est pas possible de déterminer le sens de présentation des organes (de la bouche à l'anus, ou de l'anus à la bouche ?).

Dans la littérature, la plupart des travaux concernant les ontologies et les lexiques visent à construire une ontologie en s'appuyant sur un lexique existant, ce qui est le contraire de ce que nous proposons ici. Cependant, la production de lexique textuel à partir d'ontologie a déjà été envisagée (Hirst, 2009), notamment dans des domaines techniques bien définis. Le problème que nous avons rencontré pour ordonner les éléments du lexique et les libellés dans chaque entrée est aussi rencontré par les outils générant des descriptions

en langage naturel à partir d'une ontologie, tel que ELEON/NATURALOWL (Konstantopoulos *et al.*, 2011). Ces outils génèrent une définition textuelle d'un concept d'une ontologie, à partir des relations qui ont été définies. L'ordre dans lequel les relations sont prises en compte et apparaissent dans le texte doit être configuré par l'utilisateur.

De futurs travaux porteront sur l'alignement de VCM avec des terminologies moins structurées que la SNOMED CT, telles que la CIM10, en s'appuyant sur l'alignement déjà construit, ainsi que l'alignement de l'ontologie sur une ontologie de fondement et la complétion de VCM en fonction des manques identifiés lors des alignements.

Références

Cornet R. (2009). Definitions and qualifiers in SNOMED CT. *Methods Inf Med*, **48**(2), 178–183.

Dreyfuss H. (1984). *Symbol sourcebook : An Authoritative Guide to International Graphic Symbols*. John Wiley and sons.

Ely J. W., Osheroff J. A., Ebell M. H., Chambliss M. L., Vinson D. C., Stevermer J. J. & Pifer E. A. (2002). Obstacles to answering doctors' questions about patient care with evidence : qualitative study. *BMJ*, **324**(7339), 710.

Hirst G. (2009). *Handbook on ontologies*, chapter Ontology and the Lexicon.

Horrocks I. & Sattler U. (2003). Decidability of SHIQ with complex role inclusion axioms. *Artificial Intelligence*, **160**, 2004.

Kashyap V. & Borgida A. (2003). Representing the UMLS Semantic Network in OWL. In *Proceedings of ISWC 2003 (International Semantic Web Conference)*, volume 1-16.

Konstantopoulos S., Karkaletsis V., Vogiatzis D. & Bilidas D. (2011). *Language Technology for Cultural Heritage*, volume 115-132, chapter Authoring semantic and linguistic knowledge for the dynamic generation of personalized descriptions.

Lamy J. B., Duclos C., Bar-Hen A., Ouvrard P. & Venot A. (2008). An iconic language for the graphical representation of medical concepts. *BMC Medical Informatics and Decision Making*, **8**, 16.

Lamy J. B., Duclos C. & Venot A. (2009). De l'analyse d'un corpus de texte à la conception d'une interface graphique facilitant l'accès aux connaissances sur le médicament. In *Actes des 20es Journées Francophones d'Ingénierie des Connaissances*, volume 265-276, Hammamet, Tunisie : PUG.

Lamy J. B., Soualmia L. F., Kerdelhué G., Venot A. & Duclos C. (2013a). Validating the semantics of a medical iconic language using ontological reasoning. *J Biomed Inform*, **46**(1), 56–67.

Lamy J. B., Tsopra R., Venot A. & Duclos C. (2013b). A Semi-automatic Semantic Method for Mapping SNOMED CT Concepts to VCM Icons. *Stud Health Technol Inform*, **192**, 42–6.

Ma X. & Cahier J. P. (2012). Vers un système de taguage iconique basé sur Hypertopic. In *Actes des 23e journées francophones d'ingénierie des connaissances (IC)*.

Meunier J. G. (1998). The categorial structure of iconic languages. *Theory&Psychology*, **8**(6), 805–825.

Motik B., Shearer R. & Horrocks I. (2009). Hypertableau reasoning for description logics. *Journal of Artificial Intelligence Research*, **36**, 165–228.

Rosse C. & Mejino JL V. (2003). A reference ontology for biomedical informatics : the Foundational Model of Anatomy. *J Biomed Inform*, **36**, 478–500.

L'intérêt des patrons dans la gestion des connaissances liées à la création sonore

Antoine Vincent[1], Bruno Bachimont[1], Alain Bonardi[2]

[1] HEUDIASYC UMR CNRS 7253, Université de Technologie de Compiègne, France
Antoine.Vincent@utc.fr, Bruno.Bachimont@utc.fr

[2] CICM EA 1572, Université Paris 8, France
Alain.Bonardi@ircam.fr

Résumé : La production sonore est en constante évolution, et la technologie a amplifié ce mouvement. Qu'ils soient électroniques ou numériques, les nouveaux instruments sont soumis à l'obsolescence, la préservation des œuvres passant alors par leur perpétuelle mise à jour. Celles-ci doivent être réalisées dans le respect de l'œuvre, afin que les nouvelles versions soient authentiques. Pour cela, les experts ont besoin d'accéder à des connaissances sur la manière dont a été produite l'œuvre sonore. Nous proposons un langage, à base d'ontologie et de patrons de création, permettant de gérer les connaissances et leur qualité, à destination des compositeurs et des producteurs pour les aider à préparer l'archivage des productions sonores. Le langage a été implémenté dans un méta-environnement permettant de capturer des actions de création et de générer automatiquement une modélisation d'une partie de la production sonore, permettant ainsi d'obtenir une représentation du flux de production.

Mots-clés : Design pattern, ontologie, gestion des connaissances, préservation sonore, musique technologique.

1 Introduction

La musique créée avec technologie bouscule la tradition de production sonore classique, et pose de nouvelles difficultés dans l'interprétation – et par extension dans la préservation – des œuvres créées durant le dernier siècle. Certaines œuvres sonores disparaissent après quelques années à cause de l'obsolescence technologique et du manque de représentation des connaissances nécessaires en vue de les rejouer.

Les connaissances musicologiques, pertinentes pour comprendre les œuvres et les productions sonores, sont importantes : elles permettent non plus de jouer l'œuvre, mais de la reproduire pour offrir la possibilité de la réinterpréter. Pour cela, et aider les producteurs sonores lors des inéluctables mises à jour des objets sonores, il faut leur garantir un accès à une base constituée des connaissances intelligibles pour la communauté désignée et pertinentes pour comprendre les intentionnalités originales.

La problématique abordée concerne la gestion des connaissances : nous avons besoin de capitaliser et de traiter des artefacts, qu'ils soient numériques ou non, en vue de constituer des connaissances, en cherchant le bon niveau d'abstraction entre les technologies et le besoin de compréhension, et en les filtrant pour constituer une représentation de la production de qualité.

Dans la première section de cet article, nous présenterons la problématique de gestion des connaissances à partir de la production sonore et l'objectif de préservation qui en dépend par la création d'un langage de représentation des processus de création. Dans la deuxième partie, nous présenterons notre approche complétant notre ontologie de production sonore avec les patrons, permettant de guider les modélisations afin d'obtenir des représentations de qualité. Enfin, dans la dernière section, nous étudierons la portée et l'intérêt des patrons, en abordant notamment la validation du langage ainsi créé et l'utilisation qui peut en être faite.

2 Un langage pour représenter les actes de production sonore

2.1 La production sonore actuelle

La production sonore classique repose sur un ensemble d'éléments considéré comme stable dans le temps. D'une tradition pluriséculaire, ils représentent les piliers de la reproductibilité des œuvres sonores. Ces trois éléments sont :

- la *partition*, écriture musicale permettant la reproduction de l'œuvre, étant de fait en première version l'abstraction vue par le compositeur ;
- l'*organologie*, en y incluant la facture instrumentale, qui gère donc les connaissances autour de la fabrication, du fonctionnement et de la manipulation des instruments ;
- le *conservatoire*, qui transmet à chaque génération les connaissances liées à la lecture et l'écriture musicale.

Cette stabilité apparente est efficace pour les musiques classiques, mais les évolutions technologiques perturbent ce système traditionnel : la transformation des sons à la place des notes rend la partition inutilisable. Les technologies (Donin & Feneyrou, 2013) ne permettent plus de générer une représentation universelle d'une œuvre agissant comme le guide d'interprétation, représentant une généralité de l'œuvre, et présentant ainsi les éléments indispensables permettant de jouer l'œuvre, proposant ainsi une manière de l'interpréter.

De même, les technologies propriétaires et l'augmentation exponentielle du nombre d'instruments électroniques puis des logiciels ne permettent plus de documenter une lutherie comme c'était le cas pour les instruments et la musique classiques. Une organologie des instruments modernes n'est pas possible, puisque chaque création peut donner lieu à la mise en place d'un dispositif unique et non documenté (Lemouton *et al.*, 2009).

Le conservatoire, et par extension les enseignements proposées dans certaines formations propres à l'informatique musicale, tente de proposer un ensemble de bonnes pratiques, permettant de penser aux difficultés d'interprétation des œuvres créées avec technologie. En plus de la formation classique, il est souvent présenté les difficultés de préserver une œuvre sonore menant à leur disparition seulement quelques années après leur création, et les méthodes les plus simples et actuellement efficaces permettant la reproduction sonore.

La préservation des œuvres classiques consiste donc à un devoir de mémoire (conservatoire, organologie) et d'archivage (partition). Ainsi une évolution est nécessaire pour la musique créée avec technologie : mais n'existant pas de forme de partition pour ces nouvelles productions, les techniques de préservation consistent majoritairement à travailler sur les objets technologiques pour les mettre à jour et conserver leur lisibilité (Bonardi, 2013).

2.2 Représenter pour préserver la production sonore

Dans le cas d'objet numérique, qu'ils soient sonores ou non, les méthodes de préservation visent la possibilité d'en conserver l'intelligibilité, et pour cela, il existe un ensemble de techniques régulièrement mobilisées (Vincent *et al.*, 2012b) :

- la sauvegarde, dont l'objectif est de conserver les objets du moment de création, en approchant une démarche de type muséologique ;
- la migration, qui représente un but de mise à jour de l'objet pour le modifier et le porter sur les technologies du moment ;
- l'émulation, pour simuler sur une technologie contemporaine une plus ancienne ;

– la virtualisation, qui va plus loin que l'émulation puisque l'idée est de rendre l'objet technologique indépendant d'une plateforme en particulier ;
– la description, dont l'idée est d'expliciter les connaissances permettant de remobiliser l'objet dans une forme totalement indépendante de la technologie d'origine.

Ainsi, le contexte de production devient important ; il ne suffit plus d'avoir un manuel d'exécution de l'œuvre pour la reproduire : nous passons d'une prescription de l'exécution à celle de la production. La méthode de préservation traditionnelle, portée par des normes comme *OAIS* [1] (*Open Archival Information System*), vise la conservation de l'intelligibilité de l'objet par la combinaison des techniques présentées (Ball, 2006). Mais ces portages doivent être réalisés en visant le respect de l'authenticité de l'objet.

L'authenticité d'un objet consiste à vérifier qu'il n'a pas subi d'altération dans le temps, et qu'il est bien ce qu'il prétend être (donc bien conforme face à l'objet original). Dans le cas de la musique, nous chercherons principalement à retrouver les éléments qui sont nécessaires pour réinterpréter l'œuvre en respectant les idées du compositeur.

Cette vision est bien souvent difficile à décrire, mais peut être étudiée à travers l'analyse de la production : nous pouvons comprendre certaines idées à l'aide des choix du compositeur. Il est impossible de posséder l'ensemble des intentions, mais l'accès à une partie d'entre elles nous permettrait déjà d'orienter les choix dans les besoins de mises à jour des œuvres sonores. La production est difficile à décrire, car elle dépend fortement de son contexte et des outils manipulés, et les méthodologies de production différent en fonction du compositeur ou d'une production ; les connaissances sont majoritairement tacites (Nonaka & Takeuchi, 1997).

Il nous manque un niveau des connaissances permettant de les exploiter, les abstrayant des outils numériques de production afin de les garder intelligibles. Une telle représentation abstraite serait utile pour les experts qui cherchent à conserver les œuvres sonores pour continuer de les interpréter. Mais pour créer une représentation de la production, il est nécessaire de chercher comment d'une part trouver le niveau d'abstraction pour gérer ce qui provient de la production avec des outils numériques, et d'autre part mettre en place une méthodologie qui permette de dégager ce qui est répétable et utile pour une future reproduction. Ainsi la mise en place d'une abstraction de production facilite la reproductibilité d'une œuvre ou d'un objet.

2.3 Gérer les connaissances issues des traces de production

La création d'un langage de représentation de la production permettra de gérer les connaissances potentiellement nécessaires aux experts lors des mises à jour. Ce langage ne peut être une alternative à la partition, puisque qu'il ne vise pas la gestion des connaissance permettant une nouvelle exécution, mais offrant la possibilité d'en effectuer une reproduction. Une ontologie de la production sonore offre une réponse à la première difficulté de gestion des connaissances : la définition du bon niveau d'abstraction. Ainsi, l'ontologie propose les aspects *lexicaux-syntaxiques* et *sémantiques*, base des langages.

Une ontologie de la musique, telle que *Music Ontology* (Raimond *et al.*, 2007), pourrait remplir cette tâche, mais le développement d'une ontologie de la production sonore nous permet de mettre en avant la notion du *processus de production* et de la temporalité nécessaire à la création d'une génétique de la production sonore. Ce concept de *processus* est central : un flux

1. Modèle OAIS : http ://public.ccsds.org/publications/archive/650x0m2.pdf

de production a besoin d'être représenté comme une suite d'événements et d'actions, et c'est pourquoi nous avons élaboré l'ontologie DiMPO (*Digital Music Production Ontology*). Celle-ci a été créée à partir de différents standards, comme *vCard*[2] (format d'échange de données personnelles) pour conserver toutes les informations permettant de contacter les protagonistes des productions, souvent les ultimes possesseurs des connaissances (Vincent *et al.*, 2012a).

Nous nous sommes aussi basés sur le modèle *FRBR*[3] (*Functional Requirements of Bibliographic Records*) qui est un modèle conceptuel de gestion des informations contenues dans les notices bibliographiques (O'Neill, 2002), en reprenant ses quatre niveaux :

- l'*œuvre sonore* ;
- l'œuvre possède un ensemble d'*expressions* ou de versions ;
- une expression représente des *manifestations* (performances ou mises sur support) ;
- pour chaque manifestation, il y a un ensemble de séances de travail qui s'y rattachent (ce niveau diffère de l'*item* de FRBR, puisque nous ne représentons pas que des objets physiques mais un ensemble d'objets temporels).

Nous voyons sur l'image 1 un exemple de modélisation possible de *Congruences* de Michael Jarrell créée en 1989 pour flûte midi, hautbois, ensemble et électronique (Muller, 2010).

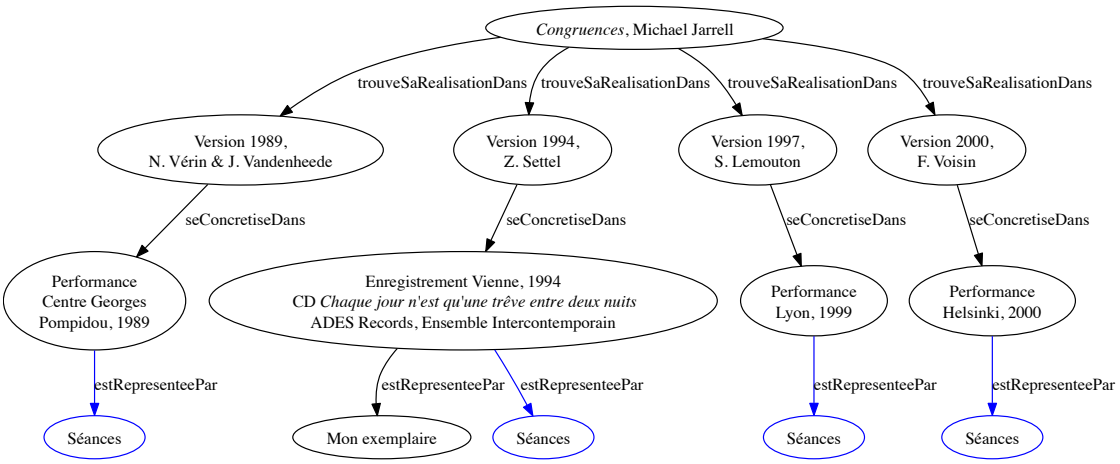


FIGURE 1 – *Congruences* possède un ensemble de versions qui représentent les évolutions technologiques qui ont été nécessaires pour reproduire l'œuvre, reproductions donnant lieu à un ensemble de manifestations.

Les concepts de l'ontologie, extraits à partir d'un ensemble de processus de production ou de reproduction étudié, avec l'aide d'experts du domaines, ont été classés sous forme d'une taxinomie en utilisant l'idée des *principes différentiels* (Bachimont, 2004). Cette méthode nous a aidé à classer des concepts provenant de productions en apparence différentes, ce qui nous a permis d'obtenir une ontologie au final avec trois niveaux (Declerck *et al.*, 2012) :

- niveau *fondationnel* : concepts présents pour articuler des concepts plus précis ;
- niveau *noyau* : concepts génériques utilisables par l'ensemble des types de productions étudiées, c'est-à-dire les œuvres créées avec technologie ;

2. Ontologie de vCard : http ://www.w3.org/TR/vcard-rdf/
3. Modèle FRBR : www.ifla.org/publications/functional-requirements-for-bibliographic-records

– niveau *domaine* : concepts qui sont manipulables par un ou plusieurs types de production, avec donc des technologies parfois développées pour une occasion particulière.

Ces différents niveaux de l'ontologie nous assurent de pouvoir utiliser l'ontologie comme base pour les modélisations, les concepts génériques étant communs à l'ensemble des productions. Mais il semble plus intéressant de toujours chercher à travailler sur les concepts les plus précis, car ils sont souvent porteurs de connaissances bien plus pertinentes. Ainsi la création d'une ontologie assez large pour couvrir un ensemble de processus de production laisse des libertés quand à son utilisation ; mais cette liberté est justement le problème : comment faire en sorte de viser une bonne utilisation de l'ontologie, afin d'obtenir des modélisations qui portent l'ensemble des connaissances pertinentes pour chaque production ?

3 La prescription des modélisations à partir d'ontologie

3.1 Difficile représentation des intentions

Les intentions du compositeur ne peuvent pas être exhaustivement et explicitement représentées, même dans le cadre de la musique classique. En revanche, il est toujours possible, avec une approche musicologique, de les étudier afin d'en chercher une explicitation. Cette étude avec la musique classique se fait à partir des partitions qui en représentent une abstraction originale. Dans le cas de la musique créée avec technologie, nous visons non plus une étude des intentions par la partition, mais par la représentation de la production, qui porte tous les choix auctoriaux.

Or, la difficulté réside dans la création de cette représentation : avec l'ontologie DiMPO, nous sommes capables de générer des modélisations des productions des œuvres sonores ; mais il faudrait pouvoir s'assurer que ces modélisations seront de qualité pour la communauté désignée, et ce selon deux points de vues :

– gestion de l'abstraction : savoir représenter correctement les connaissances pour les conserver intelligibles par la communauté ciblée, c'est-à-dire les experts musicaux ;
– guide de modélisation : filtrer les connaissances, afin de ne représenter que celles qui seront utiles afin de ne pas générer trop de bruits avec un excès de documentation.

Cette recherche de qualité s'apparente à la notion de *pragmatique* au sein d'un langage : c'est en étudiant les pratiques de la communauté, et celles du langage, que nous pouvons en améliorer son utilisation et son fonctionnement. La double gestion – modélisation et abstraction – revient en fait à rechercher, pour chaque situation de production, ce que nous devons représenter, et comment le faire. Ceci constitue notre problématique, et va plus loin que notre cadre applicatif de la création sonore : le besoin de prescrire un usage d'une ontologie afin de rechercher des modélisations de qualité permet d'améliorer et de contrôler la pertinence des travaux.

3.2 Prescrire par l'utilisation des patrons de conception

Pour compléter le langage développé à partir de l'ontologie, nous nous intéressons à cette recherche de qualité afin d'assurer la répétabilité de la production, et ainsi parvenir à l'objectif final de préservation. Pour cela, nous nous intéressons à la notion de patron de conception ou *design pattern* (Jézéquel, 2006), concept introduit en architecture par Alexander (1979) ; et utilisés sous la forme de patrons d'indexation par Isaac *et al.* (2005). Ceux-ci permettent, en les articulant avec l'ontologie, de définir l'ensemble des concepts à manipuler dans une certaine

situation de production, permettant de fait de préconiser une modélisation. Cette dernière, avec l'utilisation des patrons, sera dès lors constituée des éléments recommandés, suivant les besoins de la communauté, tout en précisant la manière de représenter les connaissances, nous assurant dès lors une utilisation optimale de l'ontologie et des expressions de qualité.

Notre approche s'inscrit dans le mouvement des *Ontology Design Patterns*[4] (ODP), mais diffère des approches dominantes : nous ne sommes pas dans une approche de conception (Gangemi & Presutti, 2009) ou de mise à jour d'une ontologie (Djedidi *et al.*, 2009), mais d'utilisation afin de développer des modèles de production d'œuvres sonores.

Les *patrons de création* (nommés ainsi en référence à la *création* d'une œuvre, c'est-à-dire sa première interprétation) forment une approche prescriptive : un patron peut être vu « comme un bloc de conception générique, l'expression d'un savoir-faire ou un guide de bonne pratique » (Fuchs *et al.*, 2010), mais il ne représente en aucun cas un ensemble de règles impératives. Ainsi le couple constitué d'une ontologie et de patrons permet d'arriver à un système laissant au final une totale liberté d'utilisation et d'application.

Pour rendre possible l'élaboration de ces patrons, nous nous appuyons sur les intentions esthétiques ou les effets sonores visés : le compositeur souhaite produire un son et le transformer, alors le patron sera là pour représenter les choix par les actions effectuées, et de manière implicite ses intentionnalités par la méthode suivie.

Nous pouvons citer comme exemple dans le cadre de la musique l'application d'un effet : en appliquant une réverbération avec un certain réglage, nous pourrons rapidement déterminer que l'effet recherché était la création d'un simple écho. C'est ainsi que la représentation des actions d'une production nous permet d'en cerner une certaine part des intentions, les patrons permettant pour chaque action d'en posséder une représentation compréhensible et complète pour en faire une analyse ultérieure.

3.3 Les patrons de création

Nos patrons de création ont été élaborés à partir de l'étude d'un ensemble de processus de production. Nous avons ainsi analysé plusieurs productions, et avec les experts du domaine, nous avons dressé des patrons qui permettent de représenter des différentes étapes des processus de production : pour une étape, l'appel au patron à chaque action effectuée permet de la documenter en ajoutant les connaissances préconisées au sein de la modélisation.

Pour prendre un exemple, étudions une chaîne de production sonore traditionnelle menant à la création d'un CD musical : différentes étapes se succèdent, que nous pourrions réduire au nombre de trois pour les processus les plus simples :

1. la *captation sonore* : acquisition des sons qui serviront de base aux travaux sonores ;
2. le *montage* : étape de transformation et d'agencements des sons ;
3. le *mastering* : préparation du CD.

Prenons l'étape de *montage* pour dresser notre premier patron : commençons par isoler, à partir de l'ontologie, tous les éléments qui seront nécessaires pour expliciter une action effectuée sur les outils technologiques, durant l'étape du montage. Les connaissant, il est nécessaire de définir l'unité permettant d'agréger toutes les informations : dans notre cas, nous débutons par

4. Ontology Design Patterns : http ://ontologydesignpatterns.org

la plus petite unité porteuse de sens parmi les concepts retenus : l'*Objet Informationnel*, celui qui va être manipulé et qui porte l'action du producteur de son.

À partir de ce concept central, nous tissons les liens qui l'unissent avec les autres concepts repérés précédemment :

- *Objet Informationnel* est Élément De *Objet Virtuel* : notre objet (sonore ou non) est utilisé dans un objet final virtuel lors d'une activité réalisée sur un outil numérique ;
- *Objet Virtuel* est Élément De *Processus* : cet objet virtuel, contenant l'ensemble des travaux sur les sons, et donc un ensemble d'objets informationnels, a été travaillé dans un cadre localisable temporellement sous la forme d'un petit intervalle de temps ;
- *Processus* compose *Manifestation* : nous retrouvons les concepts issus du modèle FRBR et permettant de gérer une classification des œuvres, de ses versions et de ses interprétations ;
- *Manifestation* concrétise *Expression* qui réalise *Œuvre*.

Ainsi, pour une étape de montage, nous avons les concepts jugés pertinent et porteurs de sens, ainsi que les relations à manipuler. Nous ajoutons alors les propriétés permettant d'affiner les connaissances portées par chaque patron. Pour cela, nous proposons de représenter les patrons en utilisant un formalisme proche de la norme UML : nous trouvons en figure 2 ce patron.

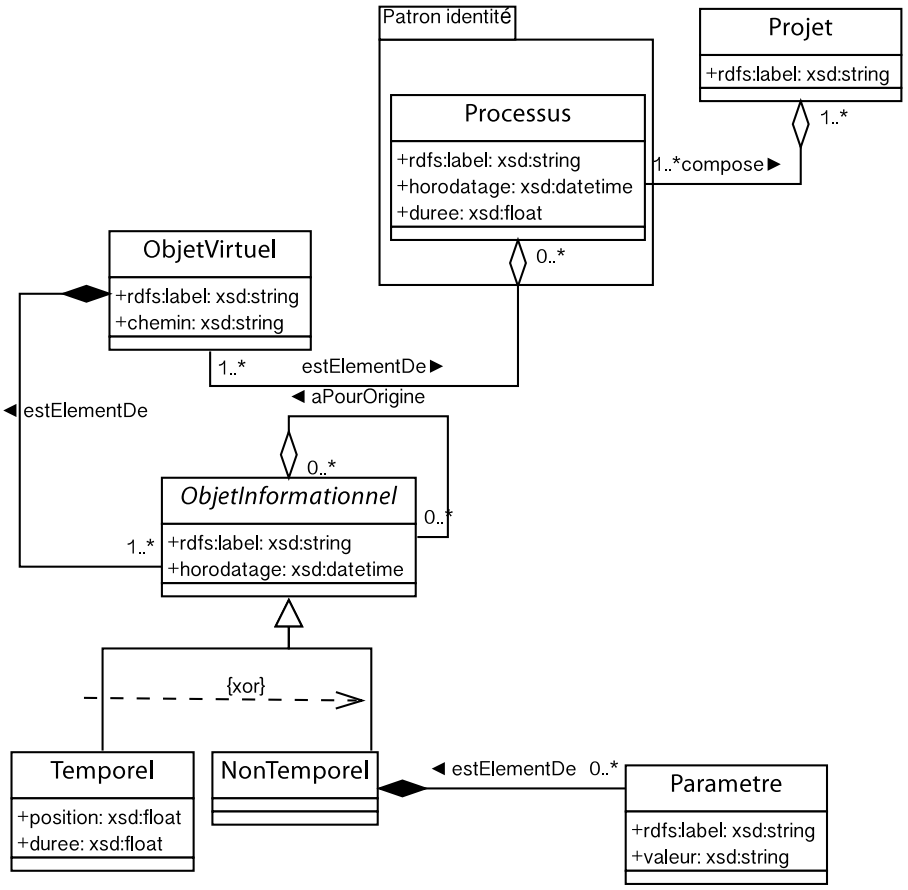


FIGURE 2 – Patron de *montage* réalisé en UML.

Ce formalisme assez simple permet de comprendre facilement comment élaborer une modélisation sans avoir besoin de maitriser parfaitement l'ontologie. Les patrons guident la modélisation en présentant tout ce qui devra être manipulé au sein de l'ontologie. Il est aussi possible de simplifier les patrons en en combinant plusieurs. Sur la figure 2, le patron de création *montage* fait référence au patron de création *identité* (figure 3), qui gère les niveaux issus de FRBR et qui préconise d'ajouter au moins un titre ; qui lui-même fait appel au patron de *contribution* (figure 3) pour inciter à toujours préciser au minimum l'identité du compositeur de l'œuvre.

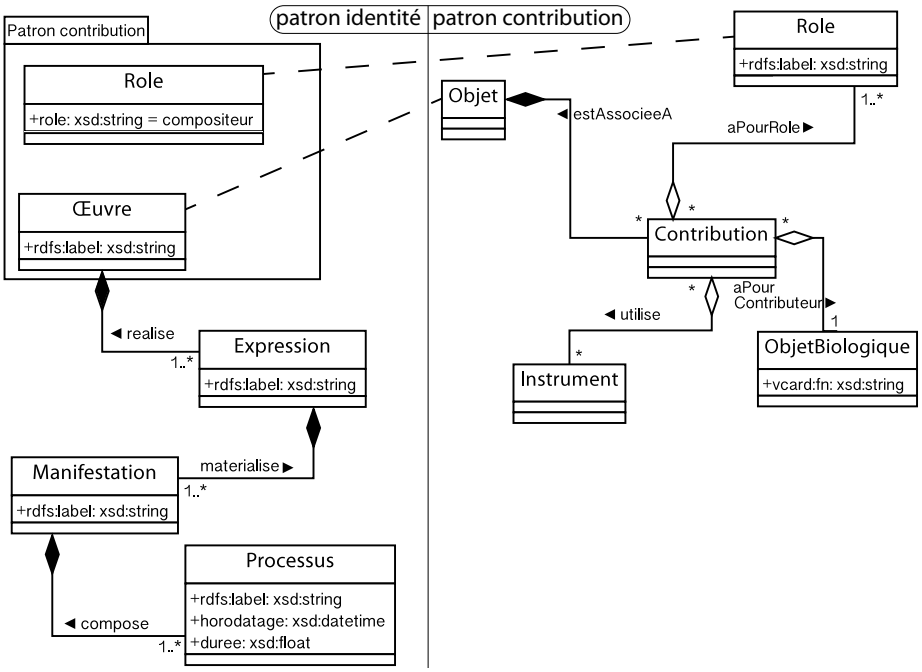


FIGURE 3 – Patrons d'*identité* et de *contribution*.

4 Les patrons pour gérer finement les connaissances à modéliser

4.1 Spécialiser les patrons

Si une production créée avec technologie suit globalement les mêmes étapes de production sonore, chaque création possèdera certaines spécificités. Cette approche de typicité nous empêche, avec nos premiers patrons, de gérer assez précisément les connaissances à représenter : elles dépendent clairement de la production, et parfois même du producteur. Nous avons besoin de préciser davantage les patrons en fonction de la production (ou de son type) pour gérer plus efficacement les connaissances et ajuster les concepts à manipuler en fonction du cas de figure : cela permet d'accroître la qualité de la modélisation, et donc la compréhension de la production pour les futurs usages de la modélisation réalisée.

L'idée est simple : l'objectif est, à partir du patron d'origine, de le préciser en changeant les concepts génériques (niveau moyen de l'ontologie) par les concepts plus précis et adaptés à la situation en cours (concepts de niveau bas de DiMPO). Prenons l'exemple en figure 4 de

la spécialisation du patron de montage pour une production réalisée sous une station audio-numérique comme Pro Tools [5], et commençons par préciser les concepts manipulés :

- l'*Objet Informationnel* central suivi est la *Piste*, cet objet étant lié à la temporalité, cela nous permet de lever le double héritage de l'objet dessiné dans le patron générique ;
- la *Piste* est Élément De *Fichier Projet* (un *Objet Virtuel*)
- le *Fichier Projet* est Élément d'une *Séance*, une unité temporelle de travail sur l'œuvre ;
- les items *Manifestation*, *Expression* et *Œuvre* ne changent pas (patron *identité*) ;
- nous ajoutons la notion d'*Élément Sonore*, un autre *Objet Informationnel*.

L'ajout de ce dernier élément n'est pas trivial : en effet, la spécialisation de patron peut en théorie être réalisée par l'expert du domaine sans avoir la parfaite maitrise de l'ontologie ; or nous observons concrètement que la déclinaison dépend pleinement des pratiques de production, et que pour obtenir une spécialisation efficace, il est nécessaire de savoir quoi représenter, si c'est possible et comment ajouter une connaissance complète.

Ce patron peut de même être représenté sous une forme proche de l'UML (figure 4), ce qui permet assez simplement de générer des modélisations ; mais dorénavant avec un patron prévu pour un certain type de production, et présentant uniquement les concepts pertinents.

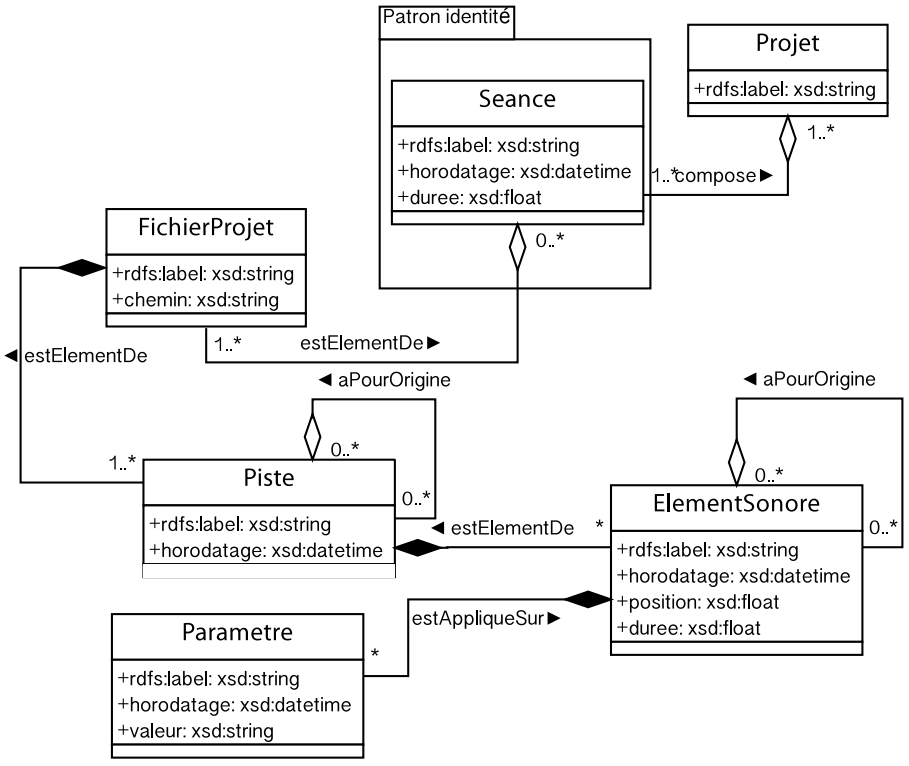


FIGURE 4 – Patron de création *montage* sous un logiciel de type Pro Tools.

5. Pro Tools est un logiciel de montage audio développé par la société Avid.

4.2 Validation pratique des patrons théoriques

Ces travaux d'élaboration de langage ont été réalisé dans le cadre du projet Gamelan [6]. L'objectif du projet consistait à construire un méta-environnement permettant d'échanger des informations avec des outils de la production sonore de manière à capitaliser le processus de création pour l'archiver, le réexploiter et l'étudier : le méta-environnement capte les *traces d'une activité instrumentée* (Prié, 2011), qu'il formate selon un modèle. Le fichier de logs ainsi élaboré permet, après application des patrons de création correspondant (implémentés dans l'environnement) d'alimenter la modélisation du processus de production : le traitement permet de passer à la constitution de connaissances exploitant dès lors uniquement les informations pertinentes.

La modélisation effectuée, il est possible d'exploiter les connaissances : sur la figure 5, nous avons l'exemple d'une requête SPARQL permettant de chercher quels ont été tous les fichiers mobilisés durant la production (les *imports* et les *exports* effectués durant le montage).

```
SELECT ?AudioFilename ?FileID ?MoveID ?Date
WHERE {  ?FileID  rdf:type             dimpo:FichierSon .
         ?FileID  dimpo:nomFichier     ?AudioFilename .
         { ?MoveID  dimpo:source         ?FileID . }  UNION
         { ?MoveID  dimpo:exporteFichier ?FileID . }
         ?MoveID  dimpo:horodatage     ?Date .
       }  ORDER BY ?Date
```

AudioFilename	FileID	MovementID	Date
"nuagesGris_extrait1_mono.wav"^^xsd:string	scenario01:FichierSon_1	scenario01:ImportAudio_1	2013-04-05T12:20:52.000
"nuagesGris_extrait2_mono.wav"^^xsd:string	scenario01:FichierSon_2	scenario01:ImportAudio_2	2013-04-05T12:20:55.000
"montage1.wav"^^xsd:string	scenario01:FichierSon_3	scenario01:ExportAudio_1	2013-04-05T12:22:14.000

FIGURE 5 – Requête d'extraction sur les fichiers mobilisés durant la production, et résultats.

La requête proposée précédemment provient du suivi du montage de l'œuvre *Nuages Gris* de Franz Liszt, interprétée par Emmanuelle Swiercz au piano pour l'album *Liszt Voyageur* [7]. Le travail de production de l'œuvre a été suivi par les logiciels du projet Gamelan, qui ont généré seuls la modélisation des opérations de montage (Vincent *et al.*, 2013).

4.3 Validation par l'usage du langage

Le langage, c'est-à-dire l'ontologie DiMPO et les patrons de création, ont ainsi été validés par l'implémentation au sein du méta-environnement du projet Gamelan, mais aussi sans passer par cet ensemble de logiciels ; l'objectif étant de nous abstraire des limitations imposées par la captation automatique et tester l'expressivité du langage autant pour les traces issues des technologies que les artefacts qui n'ont pas d'existence numérique.

La mise en place d'un scénario de validation global et objectif n'est pas aisé, car il n'y a pas d'homogénéité dans les différents processus de production des œuvres créées avec technologie.

6. Projet ANR Gamelan (2009-2013) - www.gamelan-projet.fr

7. Album *Liszt Voyageur*, Franz Liszt & Alain Bonardi (compositeurs), Emmanuelle Swiercz (interprète), label Intrada INTRA055, 2011.

Il nous semble dès lors plus logique de sélectionner des œuvres en fonction des catégories d'usages typiques. Pour ce faire, nous avons sélectionné trois profils d'utilisateur, représentatifs des pratiques, qui ont été confrontés à des situations de production et/ou de reproduction :

- un *compositeur* : une analyse d'une situation de production complexe a permis de vérifier que les connaissances utiles de son point de vue et son expérience sont représentables ;
- un *réalisateur en informatique musicale* : une création originale suivi d'une reproduction sur un autre outil permet de simuler une obsolescence technologique et vérifier que les connaissances modélisables sont suffisantes pour guider les choix de reproduction ;
- un *ingénieur du son* : en partant d'une réalisation passée, et en participant aux prémisses d'une future reproduction, nous pouvons analyser les connaissances manquantes et confirmer leur présence dans les possibilités de modélisation et les patrons de création.

Pour chaque intervenant, le scénario type avait pour objectif d'analyser les besoins en connaissances afin de les mettre en perspective avec notre langage. Les résultats confirment l'expressivité offerte mais aussi le principal point faible : nous ne pouvons pas encore nous placer en conditions réelles pour tester la qualité et l'exploitabilité des modèles, ni mesurer pleinement la pertinence des patrons de création ; en revanche nous pouvons valider les possibilités du langage et la couverture des besoins exprimées par les différents profils typiques de la communauté.

5 Conclusion

Le développement du langage a été élaborée de manière incrémentale en suivant une dizaine de productions sonores, ce qui nous permet d'avoir un langage créé en collaboration avec la communauté qui l'utilisera. L'implémentation des patrons est possible, permettant d'automatiser des modélisations, notamment les étapes réalisées sur les outils technologiques. Nous avons élaboré onze patrons génériques, certains ayant servi de base à des spécialisations, et manipulant la centaine de concepts propres à DiMPO (sans compter ceux provenant d'autres ontologies).

Cependant, il reste la difficulté de modéliser les connaissances qui n'ont pas d'existence numérique : le langage est techniquement capable de représenter des pièces documentaires, mais il est évident qu'il reste une partie des connaissances que nous ne pouvons alimenter qu'en les gérant manuellement. Nous approchons ainsi une approche d'instrumentation de la préservation, les patrons de création étant les guides pouvant accompagner cette démarche.

En revanche, nous répondons à notre besoin initial : définir une abstraction à partir de la production afin de pouvoir étudier cette dernière et cerner certaines intentionnalités. L'ontologie définit le niveau d'abstraction de pratiques qui ne sont pas stabilisées, alors que les patrons proposent un filtre séparant ce qui relève des instruments ou des intentions auctoriales : nous obtenons un système offrant une meilleure intelligibilité pour la communauté ; tout en laissant la représentation instrumentale ouverte, comme le fait la partition pour la musique classique.

Les patrons peuvent être spécialisés simplement par la précision des concepts manipulés, mais il faut souvent faire des modifications ou des ajouts pour gagner en précision. Cette approche de typicité des productions pose problème par rapport à notre hypothèse de généricité des processus, mais elle valide notre approche de validation par la sélection de profils de producteurs représentatifs des grandes pratiques pour effectuer une validation générique du langage.

Différentes perspectives peuvent être envisagées : une simplification à étudier serait de chercher à travailler à partir d'une ontologie déjà existante (nous avons cité *Music Ontology* qui pourrait être étendue). Et la spécialisation des patrons ne peut être effectuée que par les spécia-

listes de la musique et pour un ensemble d'usages : ils représentent alors une première lecture du type de production ; mais il faudrait pouvoir juger de la qualité de la spécialisation créée.

Références

ALEXANDER C. (1979). *The timeless way of building*. Oxford University Press.

BACHIMONT B. (2004). *Arts et sciences du numérique : ingénierie des connaissances et critique de la raison computationnelle.* HDR, Université de Technologie de Compiègne.

BALL A. (2006). *Briefing Paper–the OAIS Reference Model*. Citeseer.

BONARDI A. (2013). Copier/coller, ou recopier ? la transmission entre artistes des œuvres musicales avec dispositif numérique. *Technique et science informatiques*, **32**(3-4), 457–480.

DECLERCK G., AUDREY B., XAVIER A. & CHARLET J. (2012). A quoi servent les ontologies fondationnelles ? *Actes des 23èmes Journées francophones d'Ingénierie des Connaissances (IC 2012)*.

DJEDIDI R., AUFAURE M.-A. *et al.* (2009). Patrons de gestion de changements owl. In *Ingénierie des Connaissances*, p. 145–156.

DONIN N. & FENEYROU L. (2013). Symétrie recherche. In *Théories de la composition musicale au XXe siècle*, 20-21. Édition Symétrie.

FUCHS B., HUCHARD M. & NAPOLI A. (2010). Une étude sur la mise en forme de patrons de conception pour les ontologies avec l'analyse formelle de concepts. In J.-C. R. ERIC CARIOU, Ed., *Langages et Modèles à Objets (LMO)*, Langages et Modèles à Objets, p. 83–98.

GANGEMI A. & PRESUTTI V. (2009). Ontology design patterns. In *Handbook on Ontologies*, p. 221–243. Springer.

ISAAC A., BACHIMONT B. & LAUBLET P. (2005). Indexation de documents av : Ontologies, patrons de conception et d'utilisation. *16èmes journées francophones d'Ingénierie des Connaissances (IC'2005)*.

JÉZÉQUEL J.-M. (2006). Patrons de conception. *Encyclopédie Vuibert de l'informatique*.

LEMOUTON S., CIAVARELLA R. & BONARDI A. (2009). Peut-on envisager une organologie des traitements sonores temps réel, instruments virtuels de l'informatique musicale. In *Actes de la Cinquième Conférence de Musicologie Interdisciplinaire (CIM'09)*, p. 118–121.

MULLER A. (2010). La préservation et la conservation des œuvres musicales mixtes : autour du cas de congruences de michael jarrell. Master's thesis, Conservatoire National Supérieur de Musique et de Danse de Paris.

NONAKA I. & TAKEUCHI H. (1997). *La connaissance créatrice : la dynamique de l'entreprise apprenante*. De Boeck Supérieur.

O'NEILL E. T. (2002). Frbr : Functional requirements for bibliographic records. *Library resources & technical services*, **46**(4), 150–159.

PRIÉ Y. (2011). *Vers une phénoménologie des inscriptions numériques. Dynamique de l'activité et des structures informationnelles dans les systèmes d'interprétation.* HDR, Univ. Claude Bernard-Lyon I.

RAIMOND Y., ABDALLAH S. A., SANDLER M. B. & GIASSON F. (2007). The music ontology. In *ISMIR*, p. 417–422.

VINCENT A., BACHIMONT B. & BONARDI A. (2012a). Modéliser les processus de création de la musique avec dispositif numérique : représenter pour rejouer et préserver les œuvres contemporaines. In *Actes des 23ès Journées Francophones d'Ingénierie des Connaissances (IC 2012)*, p. 83–98, Paris.

VINCENT A., BACHIMONT B. & BONARDI A. (2012b). Préserver les œuvres musicales créées avec dispositif numérique par l'étude du processus compositionnel. *Les Cahiers du Numérique*, **8**(4), p. 91–118.

VINCENT A., BONARDI A. & BACHIMONT B. (2013). Étude des processus compositionnels : un langage pour représenter les processus de production sonore. In *Actes des Journées d'Informatique Musicale (JIM 2013)*, p. 9–18, Saint-Denis.

Plateforme d'interopérabilité sémantique gérant les terminologies d'interface au sein d'un espace de partage

Lamine Traore[1,2], Amina Chniti[1], Sajjad Hussain[1,2], Nicolas Griffon[2,3], Stefan Darmoni[3,2], Jean Charlet[2], Eric Sadou[2], David Ouagne[2], Eric Lepage[1,2] et Christel Daniel[1,2]

[1]AP-HP, F-75006, Paris, France; [2]INSERM, U1142, LIMICS, F-75006, Paris, France, Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1142, LIMICS, F-75006, Paris, France, Université Paris 13, Sorbonne Paris Cité, LIMICS, (UMR_S 1142), F-93430, Villetaneuse, France; [3]CISMeF, CHU Rouen, France.

Résumé : Les systèmes d'information de santé (SIS) sont complexes, hétérogènes et sont rarement interopérables, surtout au niveau sémantique. Dans le cadre du projet ANR/TeRSan, nous proposons une plateforme d'interopérabilité sémantique fondée sur les technologies du web sémantique ayant pour objectif de faciliter l'échange d'informations cliniques standardisées entre SIS au sein d'un espace de partage. L'originalité de la plateforme est d'offrir des services adaptables à tout SIS déployé dans un établissement de santé et préservant l'usage de terminologies et de modèles locaux, souvent propriétaires mais adaptés aux professionnels de santé. La plateforme repose sur 1) une modélisation ontologique des standards HL7 de partage d'information clinique, 2) la constitution d'alignements entre terminologies d'interface locales et terminologies de référence et 3) des services sémantiques permettant de transcoder des termes locaux en termes de référence en tenant compte du type de message et du contexte d'échange. Notre approche a été évaluée dans le cadre du développement d'un prototype d'échange inter-établissements d'informations cliniques dans le domaine de la télépathologie pour la demande d'avis d'expert.

Mots-clés: interopérabilité sémantique, modèle d'information, Système d'Information de Santé, ontologie, terminologie d'interface, terminologie de référence.

1 Introduction

L'interopérabilité entre systèmes d'information de santé (SIS) repose sur la standardisation de l'information clinique échangée selon des référentiels - modèles d'information (*templates*) et terminologies - partagés permettant une représentation formelle du sens des données de santé selon des standards internationaux de structuration et de codage de l'information. L'émergence de solutions opérationnelles d'interopérabilité sémantique se heurte à l'incapacité des SIS à intégrer ces référentiels tout en offrant aux professionnels de santé des interfaces de saisie d'information adaptées à leurs usages.

1.1 Problématique

1.1.1 Nécessité de l'interopérabilité sémantique en Santé

Les SIS contiennent des données cliniques hétérogènes - faits cliniques, décisions, activités - qui doivent être formalisées pour être exploitables par les machines. L'utilisation combinée de ces données formalisées et de bases de connaissances validées permet d'assister les

professionnels de santé dans leurs décisions. De plus, les données cliniques doivent être standardisées et interopérables entre établissements de santé pour pouvoir être utilisées par d'autres acteurs que leurs producteurs et dans des contextes différents. L'objectif de l'interopérabilité sémantique est de permettre le partage du sens des données cliniques entre SIS que ce soit pour l'articulation du parcours de soins, pour l'utilisation de ces données par des systèmes d'aide à la décision possiblement conçus par des tiers ou encore pour l'intégration de ces données à des entrepôts de données cliniques partagés destinés à la recherche (Mead, 2006). L'interopérabilité sémantique permet de rendre indépendant du périmètre géographique (établissement de santé, région, pays, etc.) ou du contexte (activités de soins, recherche ou la santé publique) le traitement de données.

1.1.2 Difficultés de mise en œuvre de l'interopérabilité sémantique

Malgré les efforts des organismes de standardisation dans le domaine de la santé (tels que HL7[1], CEN TC251[2] ou DICOM[3]) et malgré l'initiative internationale *Integrating the Healthcare Enterprise (IHE)* qui regroupe des professionnels de la santé précisant l'utilisation coordonnée de ces standards lors du déroulement de scénarios cliniques, les données cliniques des SIS ne sont pas nativement interopérables (European Commission, 2009). Un des obstacles à l'adoption des standards est la difficulté des éditeurs de SIS de développer et maintenir des solutions de saisie d'information structurée et codée selon ces standards qui soit adaptées à l'usage des professionnels de santé et qui s'intègrent à leur pratique quotidienne.

Ainsi, dans la pratique, les principes de structuration et de codage de l'information clinique au sein des SIS sont mis en œuvre de façon spécifique et locale au sein des établissements. Même lorsque plusieurs établissements utilisent des SIS du même éditeur, il y a très peu de partage de modèles d'information clinique d'un établissement à l'autre. Enfin, au sein d'un même établissement, les principes de structuration et de codage de l'information clinique et le niveau de granularité des informations peuvent aussi varier en fonction de la profession de santé (médecin, infirmiers, kinésithérapeutes, assistantes sociales, etc.) et au sein de ces professions, en fonction de la spécialité (cardiologie, psychiatrie, imagerie, biologie, etc.) ou encore du mode d'exercice (hospitalisation, consultation, médecine hospitalière, médecine de ville, hospitalisation à domicile, soins ambulatoire, etc.). Au total, qu'elles soient pertinentes et légitimes ou le résultat de conceptions historiques et parfois datées d'applications, les pratiques de documentation locales et les modes de représentation de l'information clinique qui y sont associés représentent des contraintes auxquelles les solutions de partage ou d'échange d'information inter-établissements doivent pouvoir s'adapter. Dès lors que l'information clinique n'est pas d'emblée interopérable lors de sa génération, à la source, des solutions d'interopérabilité sémantique sont nécessaires à la communication et au traitement de cette information au-delà du périmètre où l'information a été générée.

1.2 Hypothèse et objectif

Ce travail est réalisé dans le cadre du projet ANR/TeRSan dont l'objectif est de développer une plateforme de gestion des référentiels d'interopérabilité sémantique et de services permettant le codage de structures de données cliniques locales (*e.g.* prescriptions d'actes ou documents structurés de résultats de biologie, d'anatomie pathologique, de radiologie, etc.) selon des référentiels standard permettant leur partage inter-établissement et leur exploitation standardisée.

Notre hypothèse est que les solutions d'interopérabilité sémantique développées dans ce projet vont permettre l'échange d'information clinique standardisée entre établissements de santé tout en autorisant et préservant l'usage de modèles d'information et de terminologies

[1] http://www.hl7.org/

[2] http://www.cen.eu/CEN/Sectors/TechnicalCommitteesWorkshops/CENTechnicalCommittees/Pages/Standards.aspx?param=6232&title=CEN/TC+251

[3] http://medical.nema.org/

locales au sein des SIS de chaque établissement. Notre objectif spécifique est de valider l'approche proposée en l'intégrant à un prototype développé dans le domaine de la télépathologie. Il s'agit de spécifier et mettre en œuvre des services d'interopérabilité sémantique de sorte que des demandes d'avis émises par des pathologistes qui utilisent des SIS différents – ayant des principes locaux de structuration et de codage de l'information - soient efficacement interprétés par un destinataire qui utilise un SIS différent.

2 Contexte : plateformes d'interopérabilité sémantique de données de santé

Le partage ou l'échange de données sémantiquement annotées s'inscrit dans la problématique plus générale de la mise en correspondance de schémas (Doan, 2005). La mise en correspondance de schémas consiste à prendre deux schémas en entrée et à construire en sortie un ensemble de correspondances entre les éléments des deux schémas. Les travaux de recherche dans ce domaine ont abouti à des systèmes dont les plus significatifs sont Information Manifold (Kirk, 1995), MOMIS (Benventano, 2009), PICSEL ou encore Xyleme (Rousset, 2003). Notre travail s'inscrit dans le cadre des approches de médiation (Wiederhold, 1992). Nous nous sommes particulièrement intéressés aux travaux d'intégration de données guidés par une ontologie (Wache, 2001; Kalfoglou, 2003; Noy, 2004 ; Euzenat, 2007) et notamment à l'approche de type *global as view* dans laquelle une ontologie globale est utilisée comme source de médiation, chaque source de données alignant ses données à cette représentation pivot.

2.1 Référentiels d'interopérabilité sémantique : modèles et terminologies de réference

Dans le domaine de la santé, il existe plusieurs organismes de standardisation définissant de nombreux modèles de connaissances constituant des référentiels d'interopérabilité sémantique de l'information clinique des SIS. Parmi ces organismes de standardisation, nous distinguons les organismes tels qu'HL7, le CENTC251et DICOM qui définissent des modèles d'information de SIS (messages ou documents) et les organismes tels qu'IHTSDO ou OMS qui définissent des systèmes terminologiques de référence (terminologies, systèmes de codage ou ontologies). Les terminologies de référence sont définies par Rosenbloom et al. (Rosenbloom, 2009) comme des « terminologies conçues pour apporter une représentation complète et exacte des concepts d'un domaine donné et de leurs relations et qui sont optimisées pour la classification et la recherche de données cliniques ».

Les modèles d'information des SIS reposent sur des principes communs qui sont i) une modélisation à plusieurs niveaux d'abstraction avec la possibilité de définir des modèles spécifiques de contexte d'usage, ii) une modélisation commune des types de données de santé et iii) des règles définissant la manière d'utiliser les systèmes terminologiques (terminologies, système de codage, ontologies, etc.) lors de l'instanciation de ces modèles – propriété communément désignée par l'expression *terminology binding* (Rector, 2009).

Ainsi des modèles d'information de référence génériques (*e.g. Reference Information Model* d'HL7 version 3) peuvent être spécialisés afin de définir des modèles spécifiques de contextes d'usage (*e.g. Detailed Clinical Models* d'HL7 version 3). Si le modèle *Clinical Document Architecture (CDA)* est le modèle clinique détaillé d'HL7 version 3 le plus utilisé dans les établissements de santé, les modèles HL7 version 2 restent les standards d'échange d'information clinique les plus implémentés au monde. Les modèles d'HL7 ou du CEN TC251 intègrent des modèles communs de types de données tel que le modèle ISO 21090:2011 « Types de données harmonisées pour une interchangeabilité d'informations » standardisant la sémantique de types de données de santé (*e.g.* quantité physique, donnée codée associée à un jeu de valeurs codées et éventuellement ordonnées).

L'association entre modèle d'information et terminologies est spécifiée au niveau des « éléments de donnée » qui constituent le plus petit élément d'information au sein des modèles standards. Le standard ISO/IEC 11179-3:2013 « Registres de métadonnées » est de

plus en plus utilisé dans le domaine de la santé afin de partager des définitions non ambiguës d'« éléments de données » réutilisables faisant référence à des concepts de systèmes terminologiques. S'il est communément admis que l'ensemble des éléments de données cliniques ne peut être défini in extenso une fois pour toute et pour tous les usages par les organismes de standardisation, un certain nombre d'initiatives ont constitué des ensembles d'« éléments de données » cliniques standardisés et codés selon des terminologies de référence.

2.2 Modèles locaux et terminologies d'interface

Les établissements de santé développent le plus souvent des modèles d'information clinique locaux prenant en compte de façon évolutive les caractéristiques organisationnelles et cliniques locales. L'information clinique au sein de ces modèles est codée selon des terminologies d'interface. Les terminologies d'interface sont définies par Rosenbloom al. comme « une collection systématique de phrases (termes) du domaine de la santé définies afin de faciliter la saisie d'information par les utilisateurs au sein des SIS » (Rosenbloom, 2006). Les terminologies d'interface sont construites pour des utilisateurs et des usages spécifiques et représentent une solution de flexibilité par rapport aux problèmes d'incomplétude et de lenteur de mise à jour des terminologies de référence. Ces terminologies d'interface doivent être alignées à des terminologies de référence afin de permettre le partage et le traitement de l'information clinique (Rosenbloom, 2006; Bakhshi-Raiez, 2010; Griffon, 2012).

3 Matériel et méthode

3.1 Architecture globale de la plateforme développée dans le cadre du projet TeRSan

La plateforme d'interopérabilité sémantique est fondée sur un serveur central, des serveurs locaux situés au niveau de chaque hôpital partenaire et un ensemble de services sémantiques (voir figure 1 & figure 2).

Le serveur central gère les différentes versions de référentiels d'interopérabilité partagés (*i.e.* modèles d'information et terminologies de référence) et assure la distribution des terminologies de référence au niveau des différents serveurs locaux. Les serveurs locaux gèrent les terminologies locales et leurs alignements avec les terminologies de référence partagées.

L'interopérabilité sémantique entre SIS des hôpitaux partenaires repose sur des interfaces standardisées répondant aux différents scénarios dans le périmètre du projet (sous-traitance ou télémédecine dans les circuits de réalisation d'actes de biologie, de radiologie et d'anatomo-cyto-pathologie (ACP)). Chaque SIS des établissements partenaires doit être en capacité de développer l'interface standardisée d'échange d'information requise au sein de l'espace de partage c'est à dire d'extraire les informations à échanger et de les structurer conformément au modèle d'information standard partagé. La plateforme d'interopérabilité sémantique fournit les services sémantiques de transcodage entre terminologies d'interface et terminologies de référence permettant de transcoder l'information clinique au sein des messages ou documents échangés selon le modèle pivot partagé.

En ce qui concerne le travail présenté, les services sémantiques ont été mis en œuvre dans le cadre d'un scénario de demande d'avis d'expert en télépathologie. La figure 1 présente l'architecture des services sémantiques mise en place[4].

[4] Dans le cadre du profil d'intégration IHE Inter Laboratory Workflow (ILW) de sous-traitance d'examens de laboratoire. http://www.ihe.net/uploadedFiles/Documents/Laboratory/IHE_LAB_Suppl_ILW.pdf

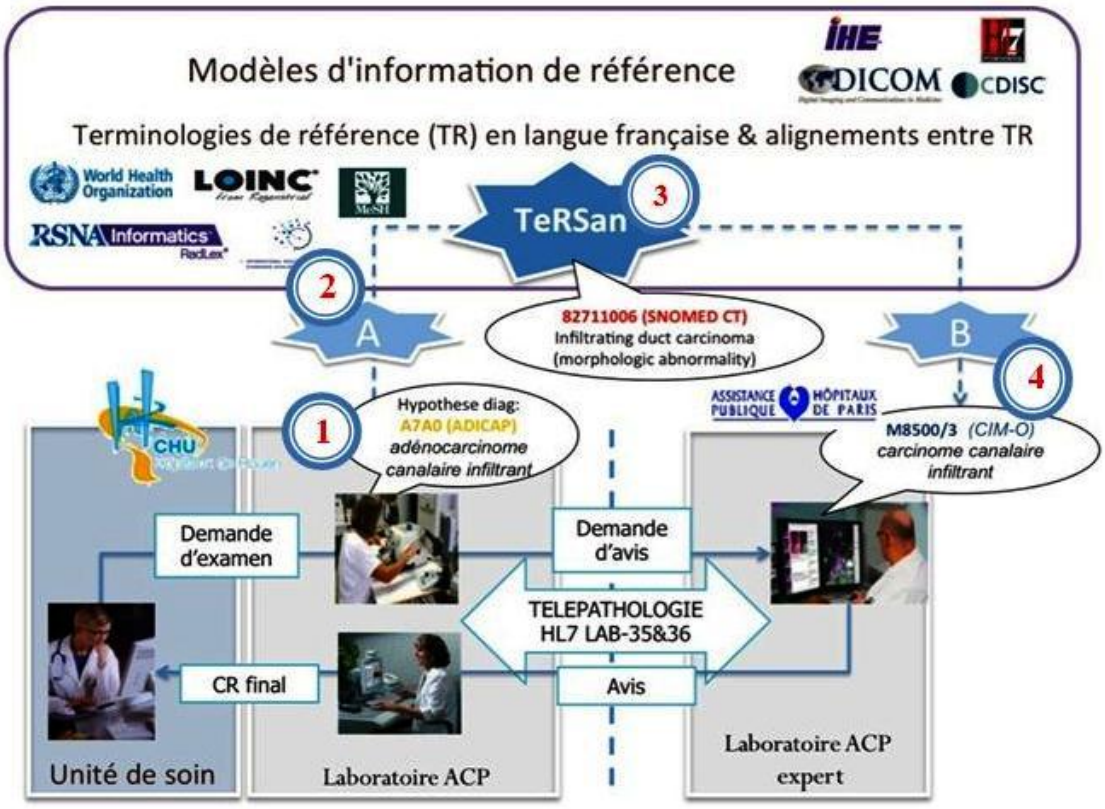


FIGURE 1– *Architecture des services d'interopérabilité sémantique TeRSan – Le serveur central (TeRSan) gère les différentes versions de référentiels d'interopérabilité partagés (i.e. modèles d'information et terminologies de référence). Les serveurs locaux (A et B) gèrent les terminologies locales et leurs alignements avec les terminologies de référence partagées. Dans l'exemple de la demande d'avis en télépathologie, les services de transcodage permettent aux établissements A et B d'échanger de l'information clinique standardisée (hypothèse diagnostique codée en SNOMED CT) tout en continuant à utiliser leurs terminologies locales (par exemple ADICAP et CIM-O).*

Lors d'un processus de télépathologie pour expertise diagnostique, le traitement de la demande d'avis comporte 4 étapes :

- Création du message : Le pathologiste du laboratoire d'ACP demandeur réalise au sein du Système de Gestion de Laboratoire (SGL) une demande d'avis d'expert concernant un examen ACP en cours. Le SGL génère un message HL7 au sein duquel les informations cliniques sont codées en utilisant la terminologie d'interface locale.
- Transcodage du message sur le site d'envoi : les services sémantiques disponibles au niveau du serveur local (A) permettent de transcoder les termes locaux du message HL7 en termes pivot des terminologies de référence du domaine.
- Echange : le message HL7 est envoyé au SGL du laboratoire ACP destinataire
- Transcodage du message sur le site de réception : les services sémantiques disponibles au niveau du serveur local (B) permettent de transcoder les termes pivot en termes locaux.

Dans l'exemple de la figure 1, la valeur de l'information clinique « Hypothèse diagnostique » est codée en ADICAP au niveau du SGL du laboratoire ACP demandeur, transcodée en SNOMED CT au niveau du serveur local (A) avant envoi puis, après réception par le laboratoire ACP expert, transcodée en CIM-O au niveau du serveur local (B) avant d'être intégrée au SGL du laboratoire expert.

3.2 Services sémantiques

Les composants de plateforme et les flux d'échange des messages/documents sont représentés dans la figure 2.

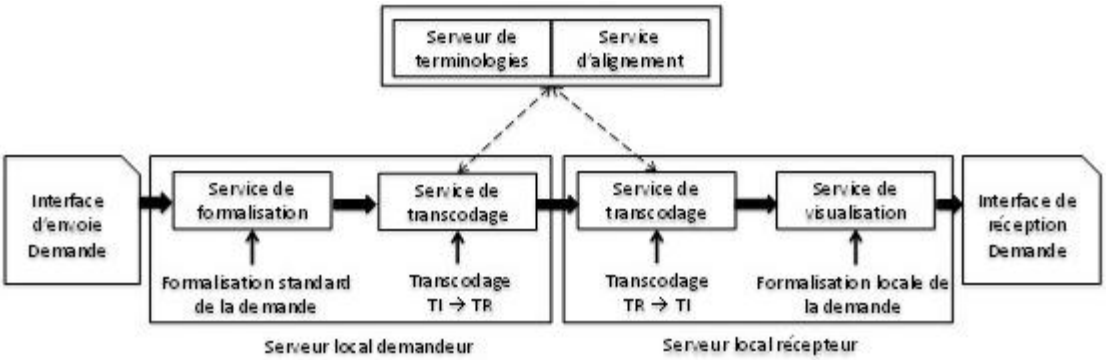


FIGURE 2 – *Composants et flux d'échange des messages/documents.*

Le service de transcodage fait appel à : i) des règles permettant l'identification des champs à transcoder (*e.g.* les champs codés dans une terminologie locale) et qui déclenchent ii) un service de recherche d'alignement qui fournit un code de la terminologie de référence pour chacun des codes de la terminologie d'interface utilisé dans le message de demande d'avis. Dans le prototype développé, afin de valider les services de transcodage, nous avons dans un premier temps développé une interface utilisateur générant une instance du message de demande d'avis ainsi qu'un service de visualisation permettant de visualiser le résultat du transcodage.

3.3 Référentiels utilisés ou construits dans le contexte de la télépathologie

3.3.1 Modèle pivot de demande d'avis ACP

Nous avons utilisé l'éditeur termApp[5] pour modéliser le modèle pivot de la demande d'avis ACP. Cet éditeur collaboratif, accessible en ligne, permet l'édition de modèles d'information de SIS conformes aux standards des organismes de standardisation HL7 ou CDISC et donc de ce fait intègre les modèles de types de données ISO 21090:2011. Cet éditeur implémente une solution d'annotation sémantique similaire à celle décrite par Rector et al. (*Code Binding Interface*) (Rector, 2009) fondée sur le modèle du standard ISO/IEC 11179-3:2013 « Registres de métadonnées (RM) »[6] (voir figure 3).

[5] http://termapp.davidouagne.com/

[6] http://www.iso.org/iso/fr/catalogue_detail.htm?csnumber=50340

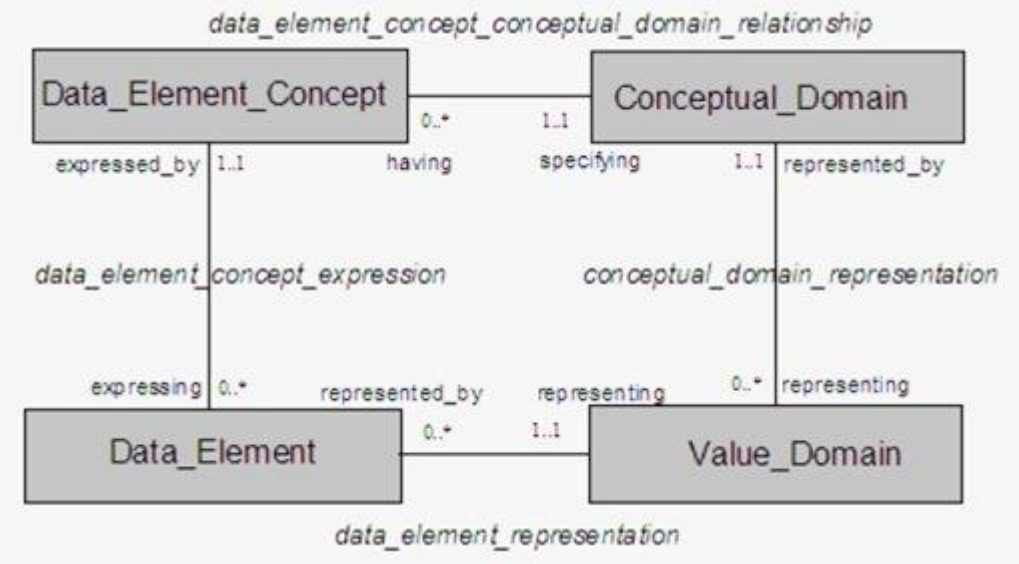


FIGURE 3 – *Modèle conceptuel du standard ISO/IEC 11179-3:2013. Chaque « élément de donnée » (Data_Element) est associé à un concept et à un domaine de valeur (Value_Domain). En cas de donnée codée (par exemple l'observation* « Hypothèse diagnostique »), *chacune des valeurs possibles du domaine de valeur (par exemple « Adénocarcinome canalaire infiltrant ») peut être explicitement associée à un code d'un système de codage (par exemple code SNOMED CT : 82711006).*

Le modèle pivot de la demande d'avis ACP a été spécifié en fonction du standard HL7[7] et enrichi afin de convenir au contexte d'usage de la demande d'avis en télépathologie. Des « éléments de données » spécifiques à ce contexte ont été modélisés. A chaque « élément de donnée » a été associé un concept médical issu d'une terminologie de référence du domaine (PathLex, LOINC, SNOMED CT) et son domaine de valeurs a été formalisé en se fondant sur le standard ISO 21090:2011. En ce qui concerne les « éléments de données » codés, chacune des valeurs possibles du domaine de valeur a été explicitement associée à un concept médical issu d'une terminologie de référence du domaine.

3.3.2 Alignement des terminologies locales/de référence

Au niveau de chaque hôpital partenaire, ont été identifiés au sein des champs du message HL7 de demande d'avis les champs correspondant à des informations cliniques codées par des terminologies d'interface et qui doivent faire l'objet de transcodage. Les terminologies d'interface utilisées au niveau de ces champs identifiés ont été extraites, modélisées selon des principes établis dans le cadre du projet TeRSan et intégrées aux serveurs locaux. Les alignements des terminologies d'interface avec les terminologies de référence ont été identifiés ou créés. Dans le cadre de la demande d'avis d'expert, les informations clés sont les hypothèses diagnostiques formulées par le pathologiste demandeur sous forme de diagnostics lésionnels. En France, selon les laboratoires d'ACP, le système de codage local utilisé pour ces diagnostics lésionnels est soit l'ADICAP[8] (1930 codes topographiques de lésions et 1648 codes lésionnels), soit la CIM-O[9] (264 codes topographiques de lésions et 1181 codes lésionnels). Des alignements ADICAP-SNOMED CT et CIM-O-SNOMED CT ont été réalisés.

[7] Message HL7 v2.5.1 OMLO21 http://www.hl7.org/implement/standards/product_brief.cfm?product_id=144, conformément à la transaction LAB-35 du profil IHE ILW

[8] Version 5.04 - Novembre 2009 (copyright ADICAP) http://www.adicap.asso.fr

[9] 3e édition – Nov 2008 (copyright OMS Genève)

4 Résultat

4.1 Modèle pivot de demande d'avis ACP

4.1.1 Modèle de la demande de sous-traitance en biologie et ACP

L'organisation hiérarchique du message HL7 utilisé dans le cadre de la transaction de sous-traitance entre laboratoires a été modélisée (voir figure 4). Ce message permet de véhiculer l'information au sein de champs organisés en segments. Les champs des différents segments contiennent les informations relatives au message lui-même, au patient, aux éléments d'assurance intervenant dans la facturation et à la demande de sous-traitance elle-même. Les informations de demande de sous-traitance sont constituées d'informations générales de la demande, d'observations cliniques pertinentes dans le contexte de la demande, d'informations relatives aux prélèvements associés à la demande et d'informations relatives à des demandes d'examens ou observations antérieures pertinentes dans le contexte de la demande de sous-traitance. Un certain nombre des champs du modèle de message sont instanciés par des informations utilisant des terminologies d'interface.

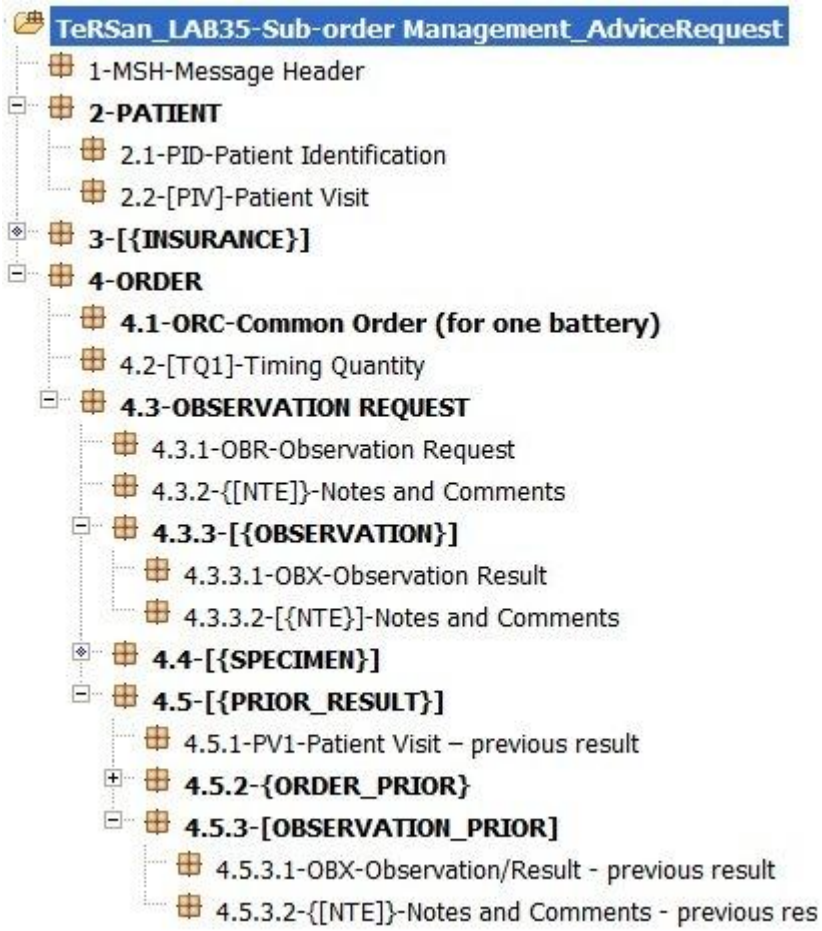


FIGURE 4– *Organisation hiérarchique du message HL7 v2.5.1 OML^O21 comportant le segment relatif au message lui-même (MSH (Message Header)), les segments relatifs au patient (PID (Patient Identification), PIV (Patient Visit), à l'assurance (Insurance), et à la demande de sous-traitance elle-même (informations générales de la demande (ORC (Common Order), TQ1 (Timing Quantity), OBR (Observation Request) et NTE (Notes and Comments)); observations cliniques pertinentes dans le contexte de la demande (OBX (Observation Results) et NTE (Notes and Comments)); informations relatives aux prélèvements associés à la demande (Specimen) et informations relatives à des demandes d'examens ou observations antérieures pertinentes dans le contexte de la demande de sous-traitance).*

4.1.2 Modélisation des observations cliniques spécifiques au contexte de la télépathologie

Une deuxième étape de modélisation nous a permis de spécialiser le modèle HL7 de demande de sous-traitance au contexte de la demande d'avis en télépathologie. Lors de cette étape, des observations spécifiques à ce contexte ont été modélisées. À chaque « élément de donnée » *Attribute Code* (OBX-3) des observations (OBX) a été associé un concept médical issu d'une terminologie de référence du domaine (LOINC ou SNOMED CT) et son domaine de valeurs (*Attribute Value* (OBX-5)) a été formalisé en se fondant sur le standard ISO 21090:2011. Lorsque la valeur de l'observation est codée (type de données de *l'Attribute Value* (OBX-5) est *Coded Element (CE)* ou *Coded With Exception (CWE)*) chacune des valeurs possibles du domaine de valeur a été explicitement associée à un concept médical issu d'une terminologie de référence du domaine (figure 5).

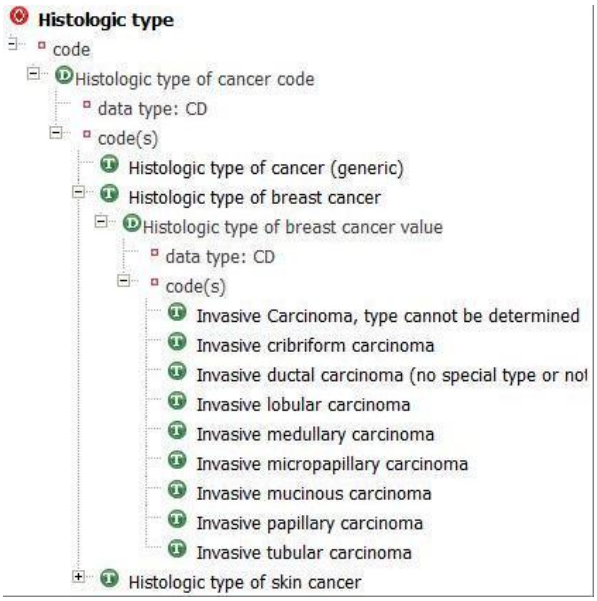


FIGURE 5 – *Modèle de l'observation Hypothèse diagnostique (type histologique). L'« élément de donnée» Attribute Code a été associé à un jeu de valeurs constitué d'un ensemble de codes LOINC correspondant à la notion de « type histologique » défini dans différents contextes (type histologique générique, type histologique de cancer du sein, type histologique de cancer cutané, etc). Pour chacun de ces codes est défini l'« élément de donnée» Attribute Value correspondant permettant d'expliciter un jeu de valeurs possibles sous la forme d'un ensemble de codes SNOMED CT.*

Un certain nombre des observations modélisées sont instanciées par des informations codées en utilisant des terminologies d'interface. Le tableau 1 présente pour trois exemples d'observations les champs du message correspondant, un exemple d'instanciation et les terminologies (locales et de référence) utilisées pour le codage de l'information.

TABLEAU 1 – *Observations spécifiques du message HL7 v2.5 de demande d'avis. Terminologies locales et de référence utilisées.*

Champ HL7 v2.5	Information	Exemple	Système de codage local	Système de codage pivot
OBX-3	Observation *(Attribute Code)*	Hypothèse diagnostique (type histologique)	TI d'observations	SNOMED ou LOINC
OBX-5	Valeur de l'observation *(Attribute Value)*	Carcinome canalaire infiltrant du sein	ADICAP ou CIM-O	SNOMED ou PathLex
OBX-3	Observation *(Attribute Code)*	Renseignements cliniques (problèmes)	TI d'observations	SNOMED ou LOINC
OBX-5	Valeur de l'observation *(Attribute Value)*	Diabète insulino-dépendant	TI locales, CIM10, CCAM	SNOMED, CIM10, CCAM
OBX-3	Observation *(Attribute Code)*	Traitement en cours	TI d'observations	SNOMED ou LOINC
OBX-5	Valeur de l'observation *(Attribute Value)*	Nolvadex	Livret thérapeutique local	ATC
OBX-3	Observation *(Attribute Code)*	CA 15.3	TI de résultat biologique	LOINC
OBX-5	Valeur de l'observation *(Attribute Value)*	40	Sans objet	Sans objet
OBX-6	Unité	U/mL	TI d'unité locale	UCUM

4.2 Services sémantiques et prototype

Le prototype implémenté permet l'envoi d'une demande d'avis d'expert en ACP entre deux établissements différents (*i.e.* APHP et CHU de Rouen). Dans ce prototype, nous nous sommes focalisés principalement sur les champs « Hypothèse diagnostique » et « Informations cliniques (problèmes, traitements en cours, résultats biologiques récents) » d'une demande d'avis. Dans notre contexte expérimental, si nous considérons l'exemple de l'information « Hypothèse diagnostique », l'AP-HP (établissement demandeur) code les valeurs possibles de cette observation en utilisant la terminologie ADICAP alors que l'établissement récepteur (CHU de Rouen) utilise la terminologie CIM-O. La terminologie de référence utilisée pour ce type d'information est la SNOMED CT. Si le champ « Hypothèse diagnostique » contient par exemple la valeur « adénocarcinome canalaire infiltrant » correspondant au code ADICAP : A7A0 à l'AP-HP, le message de demande d'avis à réception comportera, pour ce même champ, la valeur « carcinome canalaire infiltrant » correspondant au code CIM-O : M8500/3.

4.2.1 Alignement des terminologies

Au niveau de chaque établissement, les champs du message HL7 codés en utilisant des terminologies locales ont été identifiés. Les codes locaux des «éléments de donnée» et de leurs valeurs possibles ont été alignés avec les codes correspondant de la terminologie de référence. Un service de recherche d'alignement a été développé et est disponible au niveau de chaque serveur de terminologie local (voir figure 2).

4.2.2 Transcodage dynamique

Lors de l'envoi de la demande d'avis, les informations à échanger sont i) dynamiquement restructurées en correspondance avec le modèle pivot de la demande tout en utilisant les termes locaux et ii) dynamiquement transcodées grâce au service de transcodage qui permet de générer le terme de référence pour chaque terme local de la demande envoyée, (voir figure 2). Dans notre exemple nous aurons le résultat suivant :

« Hypothèse diagnostique » : A7A0^adénocarcinome canalaire infiltrant^ADICAP ➔ 82711006^infiltrating duct carcinoma^SNOMED CT

Lors de la réception, symétriquement, la demande d'avis structurée en fonction du modèle pivot utilisant les termes de référence est re-transcodée grâce au service de transcodage qui permet de générer les termes locaux de la terminologie du récepteur correspondant aux termes de référence du modèle pivot. Nous aurons ainsi le résultat suivant :

« Hypothèse diagnostique » : 82711006^infiltrating duct carcinoma^SNOMED CT ➔ M8500/3^carcinome canalaire infiltrant^ CIM-O.

5 Discussion

Notre contribution aux solutions d'interopérabilité des SIS consiste en la proposition d'une plateforme permettant la standardisation de l'information clinique échangée tout en respectant les usages des professionnels de santé qui continuent à utiliser les interfaces de saisie adaptées à leur pratique quotidienne.

En plus de la mise en place d'une infrastructure de partage au sein du périmètre d'échange – hors champ de cet article – en ce qui concerne l'interopérabilité sémantique, notre approche requiert la mise en place i) d'un serveur central permettant le partage de modèles pivots et de terminologies de référence et ii) et au sein de chacun des établissements du réseau, d'un serveur local permettant de gérer les règles de transcodage et les alignements entre terminologies d'interface locales et terminologie de référence.

Le prototype implémenté repose sur une modélisation ontologique d'un modèle d'information pivot et des services sémantiques.

L'approche proposée s'inscrit dans le cadre de la mise en œuvre de services web permettant d'enrichir sémantiquement les transactions standards entre SIS (Dogac, 2006 ; Eichelberg 2005).

Dans ce contexte, une première contribution consiste à proposer une méthode et un outil de modélisation des messages ou documents HL7 intégrant les modèles de types de données de santé ISO 21090 et une solution d'annotation sémantique de ces modèles reposant sur le standard ISO/IEC 11179-3:2013 permettant de définir la manière d'utiliser les systèmes terminologiques (terminologies, système de codage, ontologies, etc.) lors de l'instanciation de ces modèles.

Par rapport aux travaux en cours proposant des services web afin de transformer l'information clinique représentée selon des standards différents ou des versions de standards différents (Dogac, 2006), notre approche consiste à adapter ces services de sorte à ce que l'échange et l'exploitation d'une information clinique standardisée entre établissements de santé respecte l'usage de terminologies et de modèles locaux.

Le prototype implémenté a permis de valider l'approche proposée dans le contexte fonctionnel très spécifique de l'envoi d'une demande d'avis d'expert en ACP où le nombre et le type d'informations cliniques transcodées – hypothèses diagnostique, problèmes et traitement en cours – est limité.

Sur le plan méthodologique, afin de permettre la généralisation de l'approche et de la rendre plus flexible, nous allons implémenter des règles de transcodage permettant lors de l'échange de messages d'identifier, au sein d'une instance d'un modèle pivot, les informations à transcoder.

Sur le plan applicatif, nous allons étendre le périmètre fonctionnel du prototype afin de permettre la transmission des réponses aux demandes d'avis l'avis. Par ailleurs, nous devons également formaliser au sein des modèles proposés de demande d'avis et d'avis le lien entre les informations cliniques échangées et les images ACP associées à l'examen faisant l'objet de la demande d'avis. La formalisation de l'information clinique associée aux images permet à terme de concevoir et mettre à disposition des pathologistes experts au sein du réseau de télépathologie des solutions d'analyse d'images exploitant le contexte clinique.

Remerciements

Ces travaux ont été financés par l'Agence National de la Recherche, programme Technologie pour la Santé, dans le cadre du projet TeRSan (Terminologie et Référentiels d'interopérabilité en Santé) ANR-11-TECS-019. http://www.chu-rouen.fr/tersan. Nous remercions Christophe André, Sylvie Cormont, Vincent Galais, Déa Giardella, Naémé Nekooguyan, Julien Grosjean, Jean-Marie Rodrigues, Florence Amardeilh et Lydia Bascarane pour leur contribution dans le cadre du projet TeRSan et de ce travail.

Références

Bakhshi-Raiez F, Ahmadian L. *et al.* (2010). Construction of an interface terminology on SNOMED CT. Generic approach and its application in intensive care. *Methods of Information in Medicine*, 49:349-359.

Beneventano D, Sorrentino S, Orsini M, Po L. (2009). The MOMIS-STASIS approach for Ontology-based Data Integration, *1st International Workshop on Interoperability through Semantic Data and Service Integration (ISDSI 2009),* Camogli (Genova), Italy, June 25th, 2009

Doan A. & Halevy A. (2005). Semantic-integration research in the database community - A brief survey. *AI Magazine*. SPR; 26(1):83–94.

Dogac A, Laleci G, Kirbas S, Kabak Y , Sinir S, Yildiz A, Gurcan Y. (2006). Artemis: Deploying semantically enriched Web services in the healthcare domain. *Information Systems. The Semantic Web and Web Services.* Volume 31, Issues 4–5, June–July, Pages 321–339

Eichelberg M, Aden T, Riesmeier J, Dogac A, Laleci G. (2005). A Survey and Analysis of Electronic Healthcare Record Standards . *ACM Computing Surveys*, Vol. 37, No. 4, December, pp. 277–315

Euzenat J & Shvaiko P. (2007) Ontology matching. Heidelberg (DE): Springer-Verlag.

Griffon N, Savoye-Collet C, Massari P, Daniel C, Darmoni SJ. (2012). An interface terminology for medical imaging ordering purposes. *AMIA Annu Symp Proc.* 2012;2012:1237-43

Kalfoglou Y, Schorlemmer M. (2003) Ontology mapping: the state of the art. *Knowledge Engineering Review*. 18(1):1–31.

Kirk T, Levy A , Sagiv Y , Srivastava D. (1995). The Information Manifold. In *Proceedings of the AAAI 1995 Spring Symp. on Information Gathering from Heterogeneous, Distributed Enviroments*

Mead CN. (2006). Data interchange standards in healthcare IT--computable semantic interoperability: now possible but still difficult, do we really need a better mousetrap? *J Healthc Inf Manag*. Winter;20(1):71-8.

Noy N. (2004) Semantic integration: a survey of ontology-based approaches. *SIGMOD Rec*. 33(4):65–70.

Rector A, Qamar R & Marley T. (2009) Binding ontologies and coding systems to electronic health records and messages. *Applied Ontology* 4(1):51-69

Rosenbloom S., Brown S. *et al.* (2009). Using SNOMED CT to represent two interface terminologies. *JAMIA*, 16:81-8.

Rosenbloom S., Miller R. *et al*. (2006). Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. *JAMIA,* 13:277-88.

Rousset C, Reynaud C et al. (2003). Two illustrative information integration agents. Intelligent information agents research and development in europa: An AgentLink perspective, 50–78.

Wache H, Vögele T et al. (2001) Ontology-Based Integration of Information - A Survey of Existing Approaches. p. 108–17.

Wiederhold G. (1992). Mediators in the architecture of future information systems.Computer; 25(3):38–49.

Peuplement automatique d'ontologie à partir d'un catalogue de produits

Céline Alec[1], Brigitte Safar[1], Chantal Reynaud-Delaître[1], Zied Sellami[2], Uriel Berdugo[2]

[1] LRI, CNRS UMR 8623, Université Paris-Sud, France
prénom.nom@lri.fr

[2] Wepingo, 6 Cour Saint Eloi, Paris, France
prénom.nom@wepingo.com

Résumé :
Nous proposons dans cet article une approche de peuplement automatisé d'une ontologie à partir de données issues de catalogues de produits. Le peuplement automatisé est vu ici comme un problème d'annotation de documents. Dans notre contexte, les documents à annoter sont des descriptions relativement pauvres ce qui rend irréalisable un peuplement totalement automatique.

Nous proposons une approche en deux étapes : (1) une étape semi-automatique d'annotation portant sur un petit ensemble de données ; (2) une étape entièrement automatique d'annotations d'autres données basées sur des mécanismes d'apprentissage automatique exploitant les résultats de la première étape. L'originalité de ce travail consiste en une approche incrémentale de raffinement des annotations qui permet de générer des annotations même dans un contexte très restreint. Le travail décrit a été appliqué sur des jeux de données réelles concernant des jouets.

Mots-clés : Peuplement d'ontologie, Annotation sémantique, Application dans le domaine du e-commerce.

1 Introduction

Ce travail a été réalisé dans le cadre d'un partenariat entre le LRI et la startup Wepingo [1], qui développe des systèmes de recommandation de produits à des internautes. Pour faciliter la conception de systèmes adaptables à différentes catégories de produits, l'idée est de s'appuyer sur des ontologies des domaines des produits recommandés et sur les instances de produits associées aux ontologies. Dans le cadre de cette collaboration, Wepingo a mis à notre disposition une ontologie de domaine composée de concepts sans instance ainsi que des catalogues de produits de fournisseurs. Notre objectif est alors de proposer une application permettant de peupler l'ontologie à partir des éléments contenus dans ces catalogues, c'est-à-dire, de mettre en relation de façon automatisée des instances de produits avec des concepts de l'ontologie en s'appuyant sur les données textuelles décrivant ces instances. Cette mise en relation sera représentée par des annotations sur les produits, puis les produits annotés seront introduits en tant qu'instances dans l'ontologie pour être accessibles au système de recommandation.

Dans la pratique, la pauvreté relative des informations sémantiques présentes dans l'ontologie, la grande hétérogénéité des descriptions des produits des catalogues et leur manque de contextualisation rend la tâche d'annotation irréalisable de façon totalement automatique.

Notre approche consiste donc à commencer par concevoir un outil logiciel qui aide un concepteur humain à établir des liens entre des catalogues de produits et l'ontologie du domaine.

1. http://www.wepingo.com/fr-fr/

Nous avons mis en œuvre dans cet outil une démarche originale de génération et d'affinement progressif des annotations qui permet de dégager de l'information même dans un contexte très restreint. Une fois un certain nombre d'instances annotées semi-automatiquement par l'intermédiaire de l'outil, un classifieur est utilisé pour identifier automatiquement les concepts associables à de nouvelles instances.

La méthode de peuplement d'ontologie que nous proposons est a priori indépendante du domaine d'étude et du type de catalogue. Elle est adaptée au peuplement d'une ontologie comportant des classifications de produits et de caractéristiques. Des expérimentations ont été faites sur des données réelles du domaine des jouets.

Après avoir exposé le cadre de ce travail (section 2), nous ferons un rappel des travaux similaires (section 3). Nous présenterons notre approche (section 4) et nous l'évaluerons (section 5). Enfin, nous conclurons et énoncerons quelques perspectives de travail (section 6).

2 Cadre de travail

2.1 L'ontologie du monde des jouets

L'application de notre approche a porté sur le domaine des jouets. Comme support au système de recommandation, Wepingo a mis en place une ontologie des jouets (figure 1), basée sur la norme ESAR définie par des psychopédagogues (Garon *et al.*, 2002).

Cette norme identifie des catégories et des caractéristiques de jouets, en deux classifications indépendantes l'une de l'autre. Les catégories de jouets font référence au type de jouet (jeu de construction, jeu de hasard, ...) et les caractéristiques aux valeurs éducatives transmises par un jeu (concentration, dextérité, ...) ou encore ses conditions d'utilisation (jeu coopératif, associatif, ...). Un exemple de catégorie est présenté tableau 1.

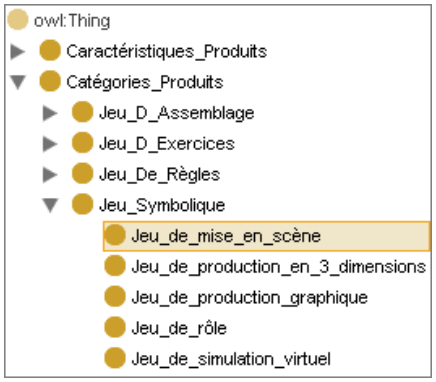


FIGURE 1 – L'ontologie ESAR

Concept $(Label)$	Jeu de mise en scène
Définition	Jeu de faire semblant dans lequel le joueur est le metteur en scène. Il réalise des scénarios élaborés dans le but de reproduire des thèmes particuliers, des scènes précises, des événements, des métiers, etc. Ces formes de jeux exigent de pouvoir mettre en scène les accessoires pertinents au contexte ou à la situation représentée.
Exemples (Ex)	playmobil, marionnette, figurine, ...

TABLEAU 1 – Le concept "Jeu de mise en scène"

L'ontologie ESAR, définie comme O_{ESAR}= (C_{ESAR}, L_{ESAR}, H_{ESAR}, Att_{ESAR}, A_{ESAR}), est limitée. C_{ESAR} est l'ensemble des concepts (33 catégories et 129 caractéristiques). L_{ESAR} est le lexique composé d'un ensemble d'entrées lexicales pour les concepts et muni d'une fonction de référence F telle que $F : 2^L \mapsto 2^C$, qui, à des ensembles d'entrées lexicales, associe des ensembles de concepts. Le lexique est composé de deux sous-ensembles de termes : $Label$ (au moins un label par concept) et Ex (exemples pour certains concepts feuilles cf tableau 1). On notera $L_{ESAR}(c)$ l'ensemble des termes de L_{ESAR} dénotant le concept c. H_{ESAR} est l'ensemble des

relations de subsomption entre les concepts. Att_{ESAR} est l'ensemble des attributs des concepts (uniquement leur définition). L'ensemble des axiomes A_{ESAR} est initialement vide. Aucune relation du domaine ne décrit les liens entre catégories et caractéristiques et ces deux classifications comportent très peu de relations de subsomption.

2.2 Les documents à annoter

Les documents (notés *Corpus*) sont des fiches décrivant un jouet par son label, sa marque, sa description (texte court non contextualisé) et sa catégorie. La catégorie ici n'est pas la même que dans l'ontologie. Elle varie beaucoup suivant le vendeur. Elle peut être très générale ("Jouet", "Jeux"), comme très spécifique ("HABA cubes et perles à assembler", "Briques"), parfois difficilement interprétable ("Bosch", "Couleurs unies"). Un exemple de descriptif de jouet est présenté figure 3a. Les formes et les contenus de ces descriptions sont très éloignés des définitions des concepts de la norme ESAR.

3 État de l'art

Annoter un document avec une ontologie consiste à rechercher dans celui-ci les fragments de texte mentionnant des concepts ou des instances de concepts appartenant à l'ontologie puis à associer ces mentions aux concepts considérés. Divers travaux d'annotation et d'extraction d'informations ont été proposés sur des domaines spécifiques. Beaucoup de ces outils, comme KIM (Popov *et al.* (2004)) ou SOFIE (Suchanek *et al.* (2009)) extraient des groupes nominaux spécifiques correspondant à des entités nommées, i.e. des noms de personnes, de lieux, d'organisations,..., repérables grâce à des grammaires formelles associées à des modèles statistiques et répertoriées dans des bases de connaissances ou des "gazeteers" (Bontcheva *et al.* (2004)).

L'identification d'instances qui ne sont pas des entités nommées est beaucoup plus délicate car aucune base ne répertorie a priori l'ensemble des instances à reconnaître et encore moins les expressions linguistiques qui leur sont associées. Ces ensembles d'instances et la terminologie propre au domaine doivent donc être recueillies pour construire la "gazeteer" adaptée à un domaine particulier. Par exemple, Amardeilh & Damljanovic (2009) prétraitent l'ensemble des termes présents dans les différentes ressources d'une ontologie (classes, instances, propriétés, valeurs de propriétés) pour en extraire un ensemble de lemmes à partir desquels est constituée la "gazeteer" associée à cette ontologie.

D'autres approches exploitent la structure du document à annoter. Par exemple, dans Amardeilh *et al.* (2005), la structure d'un document est représentée sous la forme d'un arbre conceptuel dont chaque nœud est mis en correspondance avec un concept de l'ontologie via des règles définies manuellement. De même, Aussenac-Gilles *et al.* (2013) définissent des règles d'extraction en exploitant la structure hiérarchique exprimée par les marqueurs typo-dispositionnels (police gras, italique, symbole de ponctuation ' :') au sein d'un ensemble de fiches de même format.

Les travaux cités précédemment relèvent directement du domaine de l'extraction d'informations et de l'annotation de documents. D'autres travaux, a priori plus éloignés de ces tâches, sont intéressants pour notre problématique bien qu'ils ne soient pas réalisés dans ce contexte précis. Ainsi, l'objectif de Kessler *et al.* (2012) est de vérifier l'adéquation entre des candidatures à des offres d'emploi (CV et lettres de motivation) et les offres d'emploi considérées,

c'est-à-dire évaluer la proximité entre la description d'un élément général (une offre d'emploi ou un concept d'une ontologie) et celles d'éléments plus spécifiques (des candidatures ou des instances de concept). Après avoir été soumis à différents traitements, tous les documents manipulés sont représentés par des vecteurs qui sont ensuite comparés en utilisant des combinaisons de diverses mesures de similarité (cosinus, Minkowski, ...) afin de classer les candidatures. De plus, pour être sûr de ne pas écarter trop vite une candidature, on évalue aussi sa similarité avec le vecteur représentant l'offre d'emploi enrichie des candidatures jugées pertinentes par un recruteur.

Enfin, dans Béchet *et al.* (2011), l'objectif est de peupler automatiquement une structure hiérarchique de concepts décrivant des services hôteliers, en s'appuyant sur un premier ensemble d'instances identifié par un expert. Les différents services de chaque hôtel, définis par chaque hôtelier avec son propre vocabulaire doivent être comparés aux instances initiales. Un service sera considéré comme une instance du concept correspondant à l'instance dont il est le plus proche suivant un calcul de similarité basé sur les n-grammes.

Dans notre contexte, il faut annoter des descriptions de produits comme des instances de concepts. Pour cela, il faut identifier des instances de concepts qui ne sont pas des entités nommées, au sein de documents non structurés et sans aucune homogénéité. Un certain nombre des techniques présentées précédemment sont donc complètement inadéquates. L'approche consistant à évaluer directement la proximité entre la description d'un concept et celle d'une instance est aussi inapplicable car les descriptions des concepts sont très éloignées des descriptions des produits et leur rapprochement avec des mesures de similarité ne donne aucun résultat. Bien que notre ontologie ne comporte pas initialement d'instances, l'approche proposée par Béchet *et al.* (2011) est celle qui nous est apparue comme la plus prometteuse. Pour l'identification des premières instances, nous nous sommes inspirés des travaux qui s'appuient sur des termes préalablement identifiés dans des "gazetteers" adaptées au domaine ou dans la composante terminologique d'une ontologie (Reymonet *et al.* (2007)).

4 Proposition d'une approche de peuplement d'ontologie

L'approche de peuplement de l'ontologie consiste à générer une base de connaissances $BC(O_{\text{ESAR}}, I_{\text{ESAR}}, W_{\text{ESAR}})$ à partir de l'ontologie O_{ESAR} avec $W : 2^I \mapsto 2^C$, une fonction *membre* qui, à des ensembles d'instances appartenant à I_{ESAR}, associe les ensembles de concepts de C_{ESAR} dont ils sont membres.

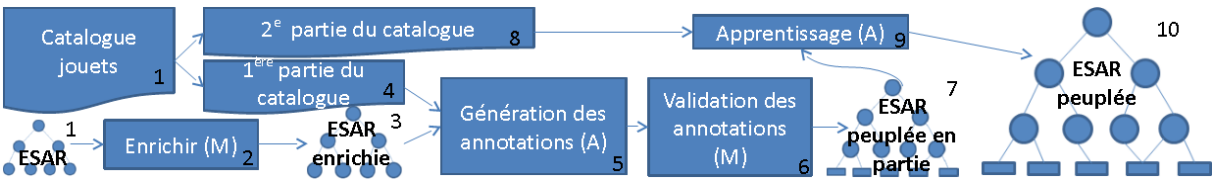


FIGURE 2 – L'approche proposée : (A) automatique, (M) manuel

Un peuplement automatique même partiel n'est possible que si l'ontologie contient les formes linguistiques associées aux concepts dont nous voulons reconnaître des instances. Notre proposition (cf figure 2) consiste donc, dans un premier temps, à enrichir (étape 1 à 3 sur la fi-

gure 2) l'ontologie de connaissances complémentaires (Section 4.1). L'ontologie enrichie est utilisée pour annoter de façon semi-automatique un échantillon de documents (étape 3 à 7 sur la figure 2). Enfin, des techniques d'apprentissage automatique exploitant ces annotations sont appliquées sur l'ensemble du corpus de documents à annoter (étape 7 à 10 sur la figure 2). L'approche d'annotation proposée comporte donc trois phases qui seront successivement décrites : la phase d'annotation d'un échantillon de documents (Section 4.2), la phase de validation des annotations de ces documents (Section 4.3), puis la phase d'annotation du corpus complet de documents (Section 4.4) basée sur l'application de techniques d'apprentissage automatique.

4.1 Enrichissement de l'ontologie ESAR

Des connaissances complémentaires nécessaires au processus d'annotation ont été ajoutées par des experts maîtrisant la norme ESAR. Ces connaissances sont de différentes natures : des termes associés aux concepts de l'ontologie (enrichissement de L_{ESAR}) et des connaissances sur ces concepts (des axiomes de A_{ESAR}).

Concernant L_{ESAR}, nous avons complété les exemples de Ex en utilisant des ressources externes. Des noms de jouets provenant d'un site internet [2] ayant utilisé la classification ESAR ainsi qu'une liste des sports provenant de Wikipedia ont été ajoutés. Nous avons par ailleurs ajouté des signes linguistiques (SL) qui sont des termes ou expressions évocateurs de concepts (par exemple "musical" et "parlant" pour le concept "Jeu sensoriel sonore") ainsi que des signes linguistiques complexes (SLcomp) de la forme "terme ET [NON] terme ET [NON] terme ..." pour aider à différencier les concepts les uns des autres. Par exemple, il existe deux types de jeux de domino : les dominos numérotés que les joueurs doivent associer (jeu d'association), et les dominos à poser debout pour construire un parcours puis à faire tomber (jeu de construction). L'utilisation de signes complexes permet de différencier ces deux jeux : le jeu de construction sera évoqué par la présence conjointe des termes "domino" et "construction" alors que le jeu d'association le sera par la présence du terme "domino" et l'absence du terme "construction". Les exemples et les signes linguistiques étant des connaissances de nature différente, nous les avons différenciés dans la représentation mais le processus d'annotation les exploite de la même façon. Après enrichissement, $L_{ESAR} = \{Label \cup Ex \cup SL \cup SLcomp\}$.

Les axiomes ajoutés dans A_{ESAR} sont exprimés sous la forme de règles propositionnelles. Il s'agit :

- d'expressions d'incompatibilités entre concepts, donnant la priorité à l'un d'eux, et de la forme "SI concept A ET concept B ALORS NON concept A" (30 règles).
- d'expressions d'inclusions de concepts de la forme "concept A IMPLIQUE concept B" (95 règles).

4.2 Annotation initiale d'un échantillon de documents représentatifs du domaine

La génération des annotations est une chaîne de traitements dont le but est de trouver un maximum d'annotations candidates exactes pour un jouet donné (catégories comme caractéristiques). Elle est composée de 3 étapes :

2. `http://www.jeuxrigole.com/liste-des-jeux.html`

1. l'établissement d'un premier ensemble d'annotations candidates qui définit le contexte d'interprétation d'un jouet ;
2. la recherche d'incohérences qui détecte au sein du contexte d'interprétation les annotations incompatibles et effectue un choix parmi elles ;
3. la complétion qui complète la liste des annotations candidates en prenant en compte des relations d'implication entre concepts.

4.2.1 Génération d'un premier ensemble d'annotations

La génération d'annotations des fiches jouets s'appuie, pour chaque concept c, sur l'ensemble $lemme(c)$ des lemmes du lexique L_{ESAR}. De même, on garde pour chaque jouet j appartenant au $Corpus$, l'ensemble $info(j)$ composé des lemmes des informations disponibles sur un jouet, i.e. son nom, sa marque, sa catégorie et sa description :

$\forall c \in C_{\text{ESAR}}, lemme(c) = lemmatisation(L_{\text{ESAR}}(c))$
$\forall j \in Corpus, info(j) = lemmatisation\{Nom(j) \cup Marque(j) \cup Cat(j) \cup Desc(j)\}$

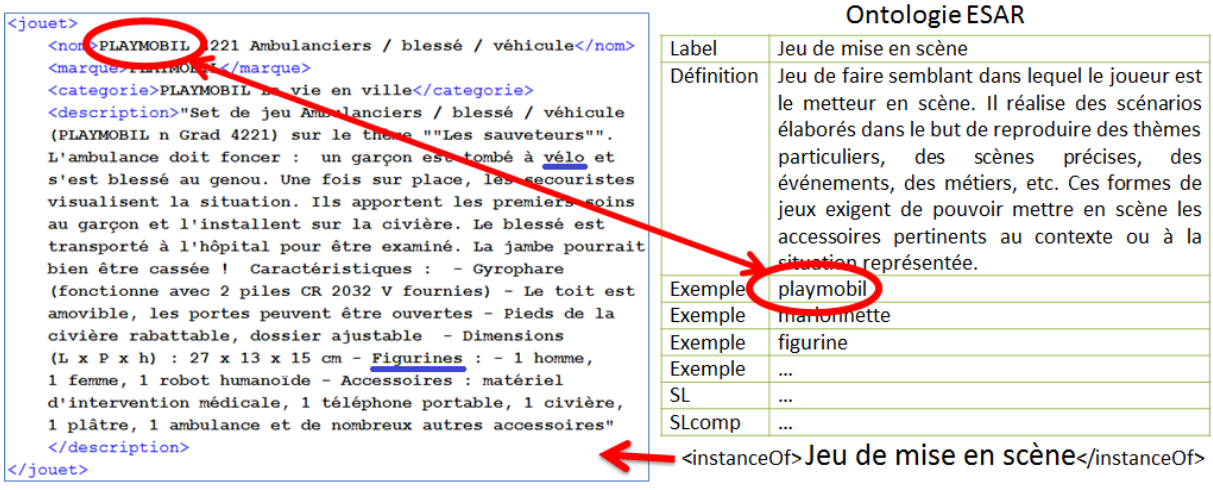
La génération des annotations est une opération de recherche d'inclusion de mots qui consiste à rechercher si un élément de $lemme(c)$ d'un concept c apparaît dans l'ensemble des informations d'un jouet j, et dans ce cas, à annoter le jouet j par le concept c (catégorie ou caractéristique) considéré :

$\forall j \in Corpus, \forall c \in C_{\text{ESAR}}$,
Si $\exists v \in lemme(c)$ tel que $v \in info(j)$ alors j `instanceOf` c.

Pour les signes linguistiques complexes, on appelle "Termes négatifs" les termes précédés de "NON" et "Termes positifs" les autres termes et on considère qu'un jouet j contient un signe linguistique complexe slc d'un concept si

$\forall\, tp \in TermesPositifs(slc), \forall\, tn \in TermesNegatifs(slc)$,
$tp \in info(j)$ et $tn \notin info(j)$

Les premières annotations produites dans cette phase définissent le contexte d'interprétation d'un jouet j, comme suit : $Ctxt(j) = \{c \mid j$ `instanceOf` $c\}$.



(a) Un exemple de descriptif de jouet (b) Le concept "Jeu de mise en scène"

FIGURE 3 – Exemple d'annotation

Par exemple, le descriptif du jouet de la figure 3 contient le terme "playmobil" qui est un "exemple" du concept "Jeu de mise en scène". Ce jouet est annoté avec le concept "Jeu de mise en scène" [3]. De même, le terme "vélo" permet de l'annoter comme "Jeu moteur" et le terme "figurines" permet de rajouter la catégorie "Jeu de mise en scène" et les caractéristiques "Créativité expressive", "Reproduction de rôles" et "Reproduction d'évènements".

Ce contexte est ensuite plus facile à analyser par les étapes suivantes que le contenu non structuré des descriptions textuelles. Pour mettre en œuvre les étapes suivantes, nous avons introduit différents ensembles de règles, chaque ensemble s'appliquant sur les résultats obtenus à la phase précédente.

4.2.2 Phase de recherche d'incohérences

La phase de recherche d'incohérences est un processus de raffinement dont le but est de détecter et d'éliminer des concepts erronés du contexte d'interprétation d'un jouet. Cette phase vise donc à améliorer la **précision** des résultats. Elle consiste à appliquer sur le contexte les règles d'incompatibilité introduites au cours de l'enrichissement. En effet, le contexte peut contenir des concepts multiples dont certains doivent être éliminés en présence d'autres. À l'issue de cette phase, on obtient un ensemble d'annotations A_1 tel que $A_1(j) \subset Ctxt(j)$.

Par exemple, le jouet de la figure 3a est annoté comme "Jeu moteur" car sa description contient le terme "vélo", alors qu'il ne s'agit pas ici d'un vrai vélo mais d'un vélo miniature associé à une figurine ("Jeu de mise en scène"). Dans ce contexte précis, l'annotation "Jeu moteur" n'est pas adaptée et il est plus facile de s'en rendre compte en la confrontant avec l'annotation "Jeu de mise en scène" également présente dans le contexte qu'en cherchant à interpréter finement la description du jouet. L'application de la règle d'incompatibilité r1 "SI Jeu de mise en scène ET Jeu moteur ALORS NON Jeu moteur" permet de supprimer l'annotation inadaptée.

4.2.3 Phase de complétion

La phase de recherche d'incohérences vise à augmenter la précision des annotations et permet à la phase de complétion de s'appuyer sur des données les plus sûres possibles. La complétion cherche à améliorer le **rappel** en exploitant toutes les inclusions entre concepts, qu'elles soient exprimées dans l'ontologie initiale ou enrichie. Elle permet d'identifier des annotations additionnelles non retrouvées lors de la phase de génération d'un premier ensemble d'annotations. À l'issue de cette phase, on obtient un ensemble d'annotations A_2 tel que $A_1(j) \subset A_2(j)$.

Par exemple, connaissant les implications "Endurance IMPLIQUE Jeu_sportif", "Jeu_sportif IMPLIQUE Jeu_moteur", un jouet annoté avec le concept "Endurance" sera par complétion également annoté par les concepts "Jeu sportif" puis "Jeu moteur".

La figure 4 montre une application des phases de recherche d'incohérences et de complétion sur l'exemple du jouet de la figure 3a. La phase de recherche d'incohérences supprime l'annota-

3. Remarquons que le fait de trouver un terme exemple d'un concept dans le nom du jouet ne suffit pas à le classer directement et définitivement comme instance du concept. Par exemple, le jouet "Playmobil pirates interactif", qui serait également annoté comme un "jeu de mise en scène", devra finalement être reconnu comme une instance de "jeu de simulation virtuelle".

tion "Jeu_moteur" en appliquant la règle r1, puis la phase de complétion ajoute les annotations "Jeu_symbolique", "Création_inventive" et "Imitation_différée".

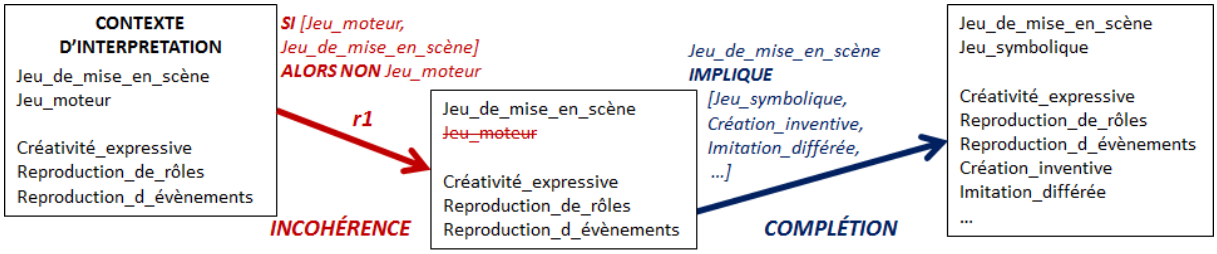


FIGURE 4 – Phases de recherche d'incohérences et de complétion sur le jouet de la fig. 3a

Ces phases s'appliquent indifféremment aux concepts catégories et caractéristiques mais dans la pratique elles ne permettent de trouver que peu d'annotations de caractéristiques car celles-ci font référence à des notions abstraites pour lesquelles les signes linguistiques sont limités. De ce fait, des étapes de raisonnement supplémentaires (cf 4.3) sont nécessaires pour déduire plus d'annotations de caractéristiques (à partir des catégories trouvées).

4.3 Validation des annotations générées

L'objectif de cette phase est de permettre à un utilisateur de confirmer ou de modifier (via une interface graphique) les annotations proposées pour un jouet et d'introduire les annotations de caractéristiques manquantes. Pour aider l'utilisateur à identifier parmi les 129 caractéristiques existantes, celles qui seront pertinentes pour le jouet considéré, le processus utilise deux heuristiques qui s'appuient sur les catégories déjà reconnues, car elles sont plus faciles à identifier.

La première heuristique consiste à identifier les caractéristiques communes aux jouets déjà traités par l'utilisateur qui sont de la même (des mêmes) catégorie(s) que le jouet à insérer. Ainsi, si un jouet est d'une catégorie A et que tous les jouets de catégorie A précédemment classés partagent un ensemble de caractéristiques E, alors l'outil propose également d'annoter ce jouet avec les caractéristiques E, en plus de celles issues du processus de génération des annotations. À l'issue de cette phase, on obtient un ensemble d'annotations A_3 tel que $A_2(j) \subset A_3(j)$. Cet ensemble A_3 est l'**ensemble des annotations proposées**.

La deuxième heuristique utilise des règles dites d'implication "potentielle" qui associent à une catégorie ses caractéristiques potentielles (i.e. qui nous paraissent être impliquées par la catégorie). Pour identifier ces règles, nous nous sommes basés sur les exemples et signes linguistiques partagés par les catégories (Cat) et les caractéristiques (Car), i.e. sur l'heuristique suivante :

$\forall\ cat_i \in Cat, \forall\ car_k \in Car,$
Si $\exists v \in lemme(cat_i)$ tel que $v \in lemme(car_k)$, alors $créer\ la\ règle : cat_i \underset{potentiellement}{\Rightarrow} car_k$.

Par exemple, "Jeu d'adresse" implique potentiellement "Coordination œil-main" car ils partagent l'exemple "toupie". 72 caractéristiques sur 129 ont été associées à au moins une catégorie et l'ensemble des règles obtenues a été ensuite complété manuellement (476 règles). Ces règles ne sont pas des règles certaines mais leur application permet d'obtenir pour un jouet j, un en-

semble supplémentaire d'annotations de caractéristiques dites annotations **suggérées**.

L'utilisateur dispose donc d'un outil muni d'une interface graphique qui, pour chaque jouet, indique des catégories et caractéristiques presque sûres (annotations proposées) et suggère des caractéristiques probables en fonction des catégories retenues (annotations suggérées). L'interface est dynamique : si l'utilisateur ajoute ou supprime des annotations, les concepts impliqués sont automatiquement ajoutés, et les suggestions de caractéristiques évoluent. Quand l'utilisateur valide, les jouets sont ajoutés dans I_{ESAR}.

4.4 Annotation du corpus complet par apprentissage basé sur l'échantillon

Après avoir décrit l'approche utilisée pour annoter un échantillon représentatif de jouets (testée sur 316 jouets), cette section présente le modèle d'apprentissage supervisé qui exploite l'échantillon pour annoter de nouveaux jouets (i.e. n'appartenant pas à l'échantillon) qui seront ajoutés à I_{ESAR}.

Nous avons utilisé le classifieur linéaire LIBLINEAR (Fan *et al.*, 2008), basé sur SVM (Cortes & Vapnik, 1995), et conseillé notamment pour la classification de documents (Hsu *et al.*, 2003). Pour chaque concept c_i, nous avons construit un classifieur SVM qui prédit pour un jouet donné si celui-ci doit être annoté par le concept c_i considéré ou pas. Nous avons donc construit 162 modèles SVM, un pour chaque concept de l'ontologie.

Pour représenter les jouets d'une manière vectorielle, nous avons testé plusieurs représentations de type sac-de-mots (Salton & McGill, 1986) : le monde est décrit avec un dictionnaire de mots et un jouet est représenté par un vecteur de la même taille que le dictionnaire de mots choisi. Chaque élément du vecteur représente un mot. Nous avons testé une représentation sac-de-mots binaire (1 pour la présence du mot dans le descriptif du jouet et 0 pour son absence) et une représentation tf-idf. Le dictionnaire utilisé est composé des lemmes des mots issus des descriptifs des jouets. Pour chaque représentation vectorielle testée, nous avons pris en compte différents sous-ensembles des attributs des jouets. Nous avons aussi appliqué une *stop-list* de mots à ne pas prendre en compte (entre autres les nombres, pronoms, prépositions, déterminants, abréviations et conjonctions) que nous appelons *stop-list* de base. Nous proposons aussi une *stop-list* plus élaborée, paramétrable par l'utilisateur, pour éventuellement ajouter d'autres catégories grammaticales à ne pas prendre en compte. Des compléments d'informations sont donnés dans la partie applicative section 5.2. La représentation vectorielle des jouets et la création des modèles SVM est entièrement automatique. Une fois les paramètres définitifs choisis, tous les jouets du catalogue sont insérés automatiquement dans I_{ESAR}.

5 Évaluation de l'approche

5.1 Évaluation du processus de génération d'annotations

Protocole expérimental. Nous ne considérons ici que les catégories de jouets car les annotations de caractéristiques sont difficiles à établir, que ce soit manuellement ou par l'outil. Pour l'évaluation, nous avons utilisé l'outil d'annotation sur un échantillon de 100 jouets construit de manière aléatoire et annotés manuellement. Les annotations proposées par l'outil ont ensuite été confrontées aux annotations manuelles.

Résultats. Le tableau 2 montre l'amélioration de la précision et du rappel apportée par les différentes étapes d'enrichissement et de raffinement. On remarque que l'amélioration la plus significative vient de l'introduction de nouveaux exemples et des signes linguistiques. Dans la confrontation des résultats, nous avons considéré comme faux un jouet annoté par plusieurs catégories dont l'une au moins était erronée. En revanche, une annotation partielle mais correcte est considérée comme juste. De ce fait, les règles de complétion ne modifient pas le résultat alors qu'en fait, elles introduisent de nombreuses annotations. Les résultats montrent que notre méthode atteint une précision satisfaisante même si le rappel est assez limité.

Étape	Précision	Rappel	F-mesure
Avant enrichissement	0,38	0,20	0,26
Exemples + signes linguistiques ajoutés	0,87	0,55	0,68
Signes linguistiques complexes	0,88	0,59	0,71
Détection incohérences (+ complétion)	0,94	0,64	0,76

TABLEAU 2 – Précision, Rappel et F-mesure du processus d'annotation

5.2 Évaluation du processus d'apprentissage automatique

Protocole expérimental. Pour évaluer la partie apprentissage de l'approche, nous nous sommes concentrés sur le concept "jeu de mise en scène". Un échantillon de jouets (316 jouets), extrait d'un catalogue particulier (Toys'R'Us) et annoté avec l'outil, constitue l'ensemble d'apprentissage. Pour l'ensemble de test, nous avons repris le même catalogue privé des jouets de l'échantillon (595 jouets) et annoté avec l'outil uniquement en terme de jeu de mise en scène ou non. Ainsi, nous construisons un modèle sur un échantillon des jouets d'un catalogue et nous observons le taux d'erreur sur les autres jouets de ce catalogue. Parmi les 36 modèles testés, nous avons opté pour celui qui génère le plus faible taux d'erreur (soit le modèle n° 12b obtenu avec les paramètres en gras italique sur le tableau 3 qui génère 2,52% d'erreur). Nous avons appliqué ces mêmes paramètres pour l'apprentissage de chacun des 162 modèles SVM créés (un par concept de l'ontologie).

Nous avons par la suite cherché à annoter (en tant que "jeu de mise en scène" ou non) avec le modèle SVM trouvé, un ensemble de 100 jouets d'un autre catalogue (Jeux et Jouets en folie) pris de manière à être le plus hétérogène possible. Soulignons que ces jouets sont très différents de ceux du premier (aucun jouet en commun). On peut donc s'attendre à ce que le modèle d'apprentissage, basé uniquement sur un ensemble représentatif du premier catalogue, ne soit pas très performant sur ces données.

Résultats. Le tableau 3 montre un extrait des pourcentages d'erreur pour le classifieur de jeux de mise en scène sur l'ensemble de test du premier catalogue. Le paramètre C du classifieur modélise le coût de violation des contraintes. Autrement dit, plus C est grand, plus on impose que les données soient sûres (non bruitées). Le descriptif correspond aux éléments considérés dans le vecteur parmi les différents attributs d'un jouet (label L, marque M, catégorie C, description D). La représentation correspond à la méthode de représentation vectorielle utilisée (binaire ou tf-idf). Les deux *stop-lists* décrites (la *stop-list* de base et celle qui supprime en plus les adverbes) ont été testées. L'ensemble d'apprentissage étant représentatif du premier catalogue tout entier, il l'est donc aussi de l'ensemble de test. Autrement dit, les jouets de l'ensemble de test sont assez proches d'au moins un jouet de l'ensemble d'apprentissage ce qui explique nos bons résultats.

Taux d'erreur					
N°	C	Descriptif	Représentation	Stop-list de base (a)	***Stop-list de base + sans adverbe (b)***
...	...	...	...	...	...
10	10	LMC	TF-IDF	6,72%	6,72%
11	10	LMCD	Binaire	3,87%	4,87%
12	***10***	***LMCD***	***TF-IDF***	3,03%	***2,52%***
13	100	LM	Binaire	9,41%	9,41%
14	100	LM	TF-IDF	9,75%	9,75%
...	...	...	...	...	...

TABLEAU 3 – Taux d'erreur d'annotation de l'ensemble de test pour les jeux de mise en scène

Le tableau 4 montre les résultats obtenus sur les 100 jouets du second catalogue avec le modèle n° 12b retenu. Parmi les 31 jouets de mise en scène, 15 ont bien été étiquetés comme tel. Aucun jouet n'a été étiqueté comme jeu de mise en scène alors qu'il ne l'était pas. On obtient donc 100% de précision et un rappel de presque 50%. Le rappel est faible car l'échantillon d'apprentissage, basé sur le premier catalogue, n'est pas représentatif des jouets du second catalogue. Cela nous semble donc très satisfaisant et nous pouvons supposer qu'en agrandissant l'échantillon d'apprentissage avec des jouets du deuxième catalogue, nous obtiendrions un meilleur rappel.

Résultats	
Taux d'erreur	16%
Précision	100%
Rappel	48,39%
F-Mesure	65,22%

TABLEAU 4 – Résultats sur 100 jouets de "Jeux et Jouets en folie"

6 Conclusion et perspectives

Dans cet article, nous avons proposé une approche originale pour associer des produits décrits dans des catalogues aux concepts d'une ontologie de domaine. Cette approche a été testé sur l'univers des jouets. Elle répond ainsi à une problématique de peuplement automatisé d'ontologie. Son originalité est d'une part la génération itérative des annotations et d'autre part la complémentarité entre les phases automatiques et semi-automatiques. Ainsi, l'approche est optimisée afin de réduire au minimum le travail de l'expert. Néanmoins le travail de celui-ci est nécessaire car la faible qualité des descriptifs de produits ne permet pas à une approche automatique d'être performante.

Les premiers résultats d'annotation des produits par leurs catégories sont prometteurs. En revanche, les caractéristiques évoquant des notions abstraites rarement utilisées dans les descriptifs, les signes linguistiques évocateurs sont plus rares et les annotations sont plus difficiles à établir.

La partie apprentissage a bien fonctionné sur les jeux de mise en scène même si ces types de jeux sont difficiles à reconnaître. Par exemple, un humain peut lire la description d'un tracteur sans comprendre s'il s'agit d'un tracteur miniature (jeu de mise en scène) ou d'un tracteur à pédales (non jeu de mise en scène). Étant donné cette difficulté pour un humain, nous estimons qu'un tel concept n'était pas simple à traiter d'une façon automatique.

L'approche exige de l'expert un travail de validation des annotations d'un échantillon représentatif de la diversité des produits. Cette tâche est manuelle et peut sembler lourde mais elle est limitée dans le temps car elle n'est à faire qu'une seule fois (modulo quelques ajustements pour prendre en compte les articles nouveaux). Le reste de l'approche est entièrement automatique.

Nous envisageons plusieurs perspectives à ce travail. Tout d'abord, trouver une solution mieux adaptée au traitement des caractéristiques. Ensuite, il faudrait utiliser une ressource externe qui permettrait d'ajouter des signes linguistiques de manière automatique et d'aider l'expert à définir les axiomes. Nous agrandirons l'échantillon afin de tenir compte des jouets de tous les catalogues. On pourrait aussi envisager d'améliorer la partie automatique en testant d'autres méthodes d'apprentissage (Bayes, Perceptron Multi-Couches, ...) et d'autres formes de représentations vectorielles (tenant compte des synonymes par exemple). Plutôt que d'utiliser un classifieur, on pourrait tester une méthode plus proche de (Kessler *et al.*, 2012), consistant à comparer un vecteur représentant plusieurs instances d'un concept donné avec un vecteur représentant un jouet à classer. Enfin, comme cette approche est indépendante du domaine et reproductible avec des connaissances adaptées, il serait intéressant de l'appliquer à d'autres domaines, tels que les cadeaux en général ou les voyages, comme souhaite le faire Wepingo.

Références

AMARDEILH F. & DAMLJANOVIC D. (2009). Du texte à la connaissance : annotation sémantique et peuplement d'ontologie appliqués à des artefacts logiciels. In *IC2009*, p. 157–168.

AMARDEILH F., LAUBLET P. & MINEL J.-L. (2005). Document annotation and ontology population from linguistic extractions. In *K-CAP '05*, p. 161–168, New York, NY, USA : ACM.

AUSSENAC-GILLES N., KAMEL M., COMPAROT C. & BUSCALDI D. (2013). Construction d'ontologies à partir de pages web structurées. In *IC2013*, p. 1–17.

BÉCHET N., AUFAURE M.-A. & LECHEVALLIER Y. (2011). Construction et peuplement de structures hiérarchiques de concepts dans le domaine du e-tourisme. In *IC2011*, p. 475–490.

BONTCHEVA K., TABLAN V., MAYNARD D. & CUNNINGHAM H. (2004). Evolving GATE to Meet New Challenges in Language Engineering. *Natural Language Engineering*, **10**(3/4), 349–373.

CORTES C. & VAPNIK V. (1995). Support-vector networks. In *Machine Learning*, p. 273–297.

FAN R.-E., CHANG K.-W., HSIEH C.-J., WANG X.-R. & LIN C.-J. (2008). LIBLINEAR : A library for large linear classification. *Journal of Machine Learning Research*, **9**, 1871–1874.

GARON D., FILION R. & CHIASSON R. (2002). *Le système ESAR : guide d'analyse, de classification et d'organisation d'une collection de jeux et jouets*. Editions ASTED.

HSU C.-W., CHANG C.-C. & LIN C.-J. (2003). *A Practical Guide to Support Vector Classification*. Rapport interne, Department of Computer Science, National Taiwan University.

KESSLER R., BÉCHET N., ROCHE M., MORENO J. M. T. & EL-BÈZE M. (2012). A hybrid approach to managing job offers and candidates. *Information Processing and Management*, **48**(6), 1124–1135.

POPOV B., KIRYAKOV A., OGNYANOFF D., MANOV D. & KIRILOV A. (2004). Kim – a semantic platform for information extraction and retrieval. *Natural Language Engineering*, **10**(3-4), 375–392.

REYMONET A., THOMAS J. & AUSSENAC-GILLES N. (2007). Modélisation de ressources terminoontologiques en owl. In *IC2007*, p. 169–181.

SALTON G. & MCGILL M. J. (1986). *Introduction to Modern Information Retrieval*. New York, NY, USA : McGraw-Hill, Inc.

SUCHANEK F. M., SOZIO M. & WEIKUM G. (2009). Sofie : a self-organizing framework for information extraction. In *WWW*, p. 631–640.

Un modèle d'annotation sémantique centré sur les utilisateurs de documents scientifiques: cas d'utilisation dans les études genre

Hélène de Ribaupierre and Gilles Falquet

ICLE, CUI, Université de Genève, Batelle, Genève, Suisse
helene.deribaupierre@unige.ch,Gilles.falquet@unige.ch

Résumé :
Lors de recherche de documents, les scientifiques ont des objectifs précis en tête. Nous avons mené des interviews auprès de scientifiques pour comprendre plus précisément comment ils recherchaient leurs informations et travaillaient avec les documents trouvés. Nous avons observé que les scientifiques recherchent leurs informations dans des éléments de discours précis, et non pas toujours dans le document en entier. A partir de cela, nous avons créé un modèle d'annotation prenant en compte ces éléments de discours. Nous avons implémenté ce modèle en OWL, et avons peuplé l'ontologie par des annotations provenant de documents dans le domaine des études genre. Nous montrons comment ce modèle permet de répondre à des requêtes précises et complexes sur un corpus de documents scientifiques.

Mots-clés : Ontologies, Annotation de documents scientifiques, Recherche d'information précise

1 Introduction

(Hannay, 2010) a écrit que les scientifiques ont de meilleurs outils pour gérer leurs données personnelles (photos et vidéo) que pour gérer ou chercher dans leurs données professionnelles. Ce constat est toujours valable, il est toujours difficile pour un scientifique de trouver le ou les bons documents qui correspondent effectivement à son besoin d'information. De plus, le nombre de documents scientifiques publiés chaque année est de plus en plus important (le nombre de documents dans Medline augmente de 0.5 millions par année (Novácek *et al.*, 2010)). Les moteurs de recherche académiques de type Google Scholar, DBLP ou Web of Knowledge indexent les documents par métadonnées et par les mots contenus dans le texte. Ils sont inefficaces dans le cas de requêtes complexes et précises telle que : « trouver tous les résultats de recherches qui utilisent une méthodologie quantitative et qui montrent que les filles sont meilleures dans les tâches de lecture que les garçons ». Pour traiter ce genre de requête, il faut entre autres être capable de détecter si les concepts cherchés apparaissent dans une partie du texte présentant un résultat de recherche ou une méthodologie. D'où le besoin d'indexer sémantiquement non seulement les mots des textes, mais également de caractériser la fonction de chaque fragment de texte (hypothèse, méthodologie, résultat, etc.). Dans cet article, nous proposons un modèle et un système d'annotation de documents scientifiques générique, prenant en compte les besoins des scientifiques. Nous avons choisi le domaine des études genre comme cas d'étude pour tester notre système, car les documents y sont très hétérogènes, allant d'études très empiriques à des documents de type philosophique, et n'utilisent que rarement le modèle structurel IMRaD (introduction, méthodologie, résultat et discussion).

2 Modèle d'utilisateur

Parmi les travaux qui étudient le comportement de recherche d'information et de lecture des scientifiques, (Bishop, 1999), en indexant des composants spécifiques dans une bibliothèque numérique (figures, conclusions, références, titres, titre de figures/tableaux, auteurs, etc.), montre que les scientifiques les utilisent pour faire des recherches d'information plus pertinentes. (Renear & Palmer, 2009), montrent que les scientifiques lisent et extraient des informations spécifiques telles que les "findings" [1], les équations, les protocoles de recherches et les données.

Pour comprendre ces besoins, nous avons mené deux études, l'une quantitative, l'autre qualitative, auprès de scientifiques de différentes communautés (de Ribaupierre & Falquet, 2011). Ces entretiens nous ont aussi permis de construire des cas d'utilisation génériques à partir des questions qu'ils se posaient avant d'utiliser un moteur de recherche et de convertir leurs questions en mots-clés. Nous avons extrait une douzaine de cas d'utilisation, dont trois exemples sont présentés ci-dessous.

Exemples de questions que les interviewés se posent	**Cas d'utilisation extrait**
Trouver les définitions de la notion d'homogénéité sémantique et si cela se calcule.	Trouver les différentes définitions d'un terme, et leurs différentes facettes [2]
Est-ce que Christine Delphy se dispute avec Patricia Roux dans un article ?	Trouver les auteurs qui ne sont pas d'accord avec l'auteur X, ou inversement, trouver les auteurs qui sont en accord avec l'auteur X
Trouver tous les auteurs qui travaillent sur la variabilité intra-individuelle du point de vue du comportement	Trouver les auteurs dans mon domaine de recherche

Nous avons aussi trouvé, que les scientifiques se concentrent sur certaines parties des documents qu'ils lisent. Les cinq types d'information que les scientifiques regardent en priorité sont (en dehors du résumé), les findings, les méthodologies, les hypothèses, les définitions et les travaux référencés (background). En entretien, nous avons confirmé l'hypothèse selon laquelle ces types ne correspondent pas forcément (même si souvent appelés de la même manière) aux parties structurelles des documents, mais bien à des fragments décrivant un de ces types pouvant se trouver n'importe où dans le document. Ainsi quand un scientifique parle de findings, il ne se réfère pas forcément à la partie structurelle portant le nom "findings", mais à tous les findings contenus dans le document.

3 Modèle d'annotation d'articles scientifiques

Il existe un certain nombre de modèles d'annotation pour les documents scientifiques. Certains auteurs (Groza *et al.*, 2007; Harmsze, 2000; Teufel & Moens, 2002; Ibekwe-Sanjuan *et al.*,

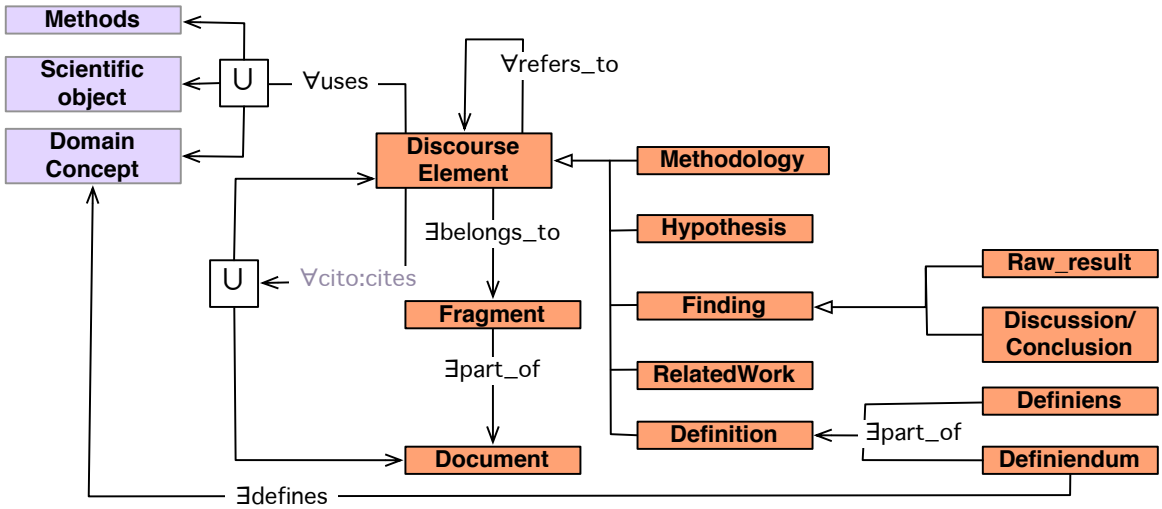
1. Il est difficile de trouver une traduction satisfaisante du mot "findings" qui regroupe à la fois les résultats bruts, les différentes trouvailles, observations discussions et conclusions du document scientifique

2. Notons à propos de cette requête que l'emploi de l'option "define" de Google est loin d'être satisfaisant. En effet, Google ira chercher des définitions provenant de glossaires connus ou de Wikipédia, donc "grand public" ou consensuelles, alors que le but de cette requête, pour le scientifique, est au contraire de trouver les définitions d'un terme proposées par des scientifiques dans un corpus de documents donné.

2008; de Waard *et al.*, 2009) proposent d'utiliser les structures rhétoriques ou les éléments de discours pour annoter les documents de manière à produire de meilleurs systèmes de recherche d'information ou pour créer des résumeurs automatiques. Ces travaux utilisent généralement des documents provenant des sciences dites "dures" telles que la biologie, la médecine ou la physique, où les documents sont très structurés, et où l'écriture des résultats, hypothèses ou méthodes est souvent formelle. Par ailleurs, seuls (Djioua & Descles, 2007), prennent en compte la définition comme type d'élément de discours. De plus, à notre connaissance, seul (Teufel & Moens, 2002) et (Djioua & Descles, 2007), annotent automatiquement les documents, les autres modèles sont utilisés pour de l'annotation manuelle. La construction de notre modèle d'annotation (voir figure 1) est essentiellement basée sur l'analyse des entretiens et du questionnaire, mais elle cherche aussi à agréger certains des modèles proposés.

Notre modèle d'annotation est basé sur trois axes : les éléments de discours, les relations explicites entre les documents et les métadonnées.

FIGURE 1 – Modèle d'annotation de document scientifique (Methods, Scientific object et Domain concept sont importés d'autres ontologies)



Eléments de discours : Ces éléments sont définis par l'ontologie SciDocAnnotation. Le contenu des éléments de discours est indexé sémantiquement à l'aide de concepts provenant d'ontologies auxiliaires : 1) ontologie(s) du domaine étudié ; 2) ontologie(s) des objets scientifiques (équations, modèles, algorithmes, théorème, etc.) ; 3) ontologie(s) des méthodes (types de méthodes, types de variables, outils utilisés, etc.). Il y a cinq types d'éléments de discours

Une Definition se compose d'un Definiens (la phrase définitoire) et d'un Definiendum (le terme défini). Le définiendum est relié par la relation defines à un concept du domaine, ainsi les différentes définitions d'un même définiendum utilisent le même concept du domaine.

Un Finding regroupe toutes les trouvailles, observations, discussions et conclusions d'un document. Le Raw_result qui défini des résultats pas encore analysés ou discutés. Alors que la Discussion/Conclusion, comme son nom l'indique, définit des résultats analysés et discutés.

Un élément de type Methodology décrit les différentes méthodes et étapes utilisées dans la recherche.

Un élément de type Hypothesis est une proposition de réponse à une question posée.

Un élément, quel que soit son type, est également de type RelatedWork s'il provient d'autres travaux. Nous sommes partis de l'hypothèse que les annotations des documents scientifiques doivent se faire sur une base de connaissance "universelle", et non pas centrée sur l'auteur. Nos études ont montré que ce ne sont pas forcément les écrits d'un auteur précis qui intéressent les scientifiques, mais une connaissance plus "universelle", qui doit par la suite être réattribuée à son auteur. Par exemple, quand une personne interrogée a répondu : « Trouver les différents articles qui traitent de l'évaluation des simulateurs chirurgicaux », cette scientifique, dans un premier temps, est intéressée à trouver tous les documents traitant de ce sujet, et cela indépendamment de l'auteur.

Références explicites d'un document/élément de discours à un autre document/élément de discours : Nous utilisons pour cela l'ontologie CiTO [3] de (Shotton, 2009). Contrairement aux moteurs de recherche académiques cités ci-dessus, nous annotons les relations, non pas au niveau document/document, mais au niveau élément de discours/élément de discours. Cela permet de résoudre des questions précises telle que « tous les findings démontrant une différence de sexe à l'école en mathématique et référençant un *résultat* de *Zazzo* ». Il devient également possible d'effectuer des analyses plus fines du réseau des citations, en fonction des types d'éléments cités ou citants.

Métadonnées : Il s'agit des données bibliographiques usuelles, telles que le nom des auteurs, le titre de l'article, le nom du journal ou de l'éditeur, etc.

4 Implémentation et Evaluation du modèle sur un cas dans le domaine des études genre

Nous n'avons pas trouvé d'ontologie du domaine des études genre, ni des objets scientifiques, ni des méthodologies que nous aurions pu importer dans notre modèle. Nous avons donc créé ces ontologies (contenant respectivement 465, 19 et 36 classes) [4].

Nous avons utilisé GATE [5], ANNIE [6], les règles JAPE et les modules de gestion d'ontologies, pour annoter automatiquement les différents éléments de discours contenus dans les documents et les concepts décrivant le contenu. Dans le but d'automatiser l'annotation des documents, nous avons défini l'élément de discours au niveau de la phrase, et le fragment au niveau du paragraphe. Après analyse manuelle d'un corpus de 20 documents, nous avons analysé les motifs syntaxiques des types d'éléments de discours à l'aide d'ANNIE et créé des règles JAPE (20 règles pour les findings, 34 règles pour les définitions, 11 règles pour les hypothèses et 19 règles pour les méthodologies).

Pour tester la qualité de l'annotation automatique nous avons annoté manuellement 555 phrases en études genre [7], créant ainsi un "golden standard". Nous avons effectué des mesures de précision/rappel sur ces phrases (voir tableau 1) qui montrent une bonne précision, mais un rappel peu élevé.

3. http ://purl.org/spar/cito

4. disponible sous http ://cui.unige.ch/~deribauh/Ontologies/

5. http ://gate.ac.uk/

6. http ://gate.ac.uk/ie/annie.html

7. les éléments de discours annotés manuellement, ne viennent pas des mêmes documents sur lesquels nous avons fait nos analyses syntaxiques.

TABLE 1 – Mesures de précision/rappel

Types d'éléments de discours	Nb de phrases	Prec.	Rappel	F1.0s
findings	168	0.82	0.39	0.53
hypothèses	104	0.62	0.29	0.39
définitions	111	0.80	0.32	0.46
méthodologies	172	0.83	0.46	0.59

Nous avons ensuite annoté automatiquement 903 articles en anglais, venant de différents journaux (étude genre et sociologie). Nous avons importé ces annotations dans un triplestore Allegrograph [8]. Nous sommes arrivés, après nettoyage, à 73'994 fragments (paragraphe) et 342'425 éléments de discours (phrases) se répartissant en : 304'747 qui n'ont aucun type, 15'449 findings, 11'813 méthodologies, 7'244 hypothèses, 3'172 définitions, parmi lesquels 2'780 travaux référencés dont 1'444 findings, 792 méthodologies, 351 hypothèses et 193 définitions. Nous avons annoté toutes les phrases sans type exportées de GATE par un élément *SentenceNotDefined*. Nous utilisons cet élément dans des heuristiques nous permettant de pallier le taux de rappel (voir ci-dessus). En effet, si un fragment contient des phrases non définies et plus de trois phrases de même type et uniquement de ce type, nous inférons que les phrases non définies sont du même type que les phrases définies. Avec ces règles, nous avons pu identifier 341 findings, 130 méthodologies, 29 hypothèses et 6 définitions supplémentaires.

Pour effectuer des tests comparatifs avec des utilisateurs nous avons défini deux interfaces de recherche sur cette base d'annotations : une interface de recherche classique par mots clés (avec un modèle de pondération TF*IDF) et une interface à facettes basée sur notre modèle (les facettes correspondant aux types d'éléments de discours).

5 Discussion / Conclusion

La principale contribution de cet article est la catégorisation et l'annotation automatique des éléments de discours basé sur un modèle d'annotation construit à partir d'interviews et de questionnaires soumis à des scientifiques. On peut considérer que le modèle est réaliste dans la mesure où une annotation automatique des documents est possible avec des outils classiques de traitement de la langue naturelle. Si le taux de rappel est bas, la précision des annotations est bonne, signifiant que si un utilisateur lance une requête dans notre système, il a une bonne probabilité de trouver une information précise, contrairement aux moteurs de recherche académiques actuels. De plus, les exemples de questions précises que les scientifiques nous ont fournis constituent une base de cas d'utilisation et de requêtes pour valider notre modèle et évaluer notre système.

Divers tests avec des utilisateurs sont en cours pour comparer notre modèle aux modèles de recherche par mots clés. Nous ne disposons pas encore d'une analyse complète de ces tests, mais les premières mesures tendent bien à montrer que notre modèle est nettement supérieur en précision. À titre d'exemple, la requête «Trouver les définitions qui contiennent le mot *gender*»,

8. http ://www.franz.com/agraph/allegrograph/

livre effectivement 143 définitions contenant le mot *gender*, alors qu'une recherche avec le mot clé "gender" fournit 13'210 éléments de discours, ce qui est de peu d'utilité.

Pour permettre l'automatisation, nous avons dû simplifier certaines relations du modèle de départ. Dans le cas des définitions, nous annotons les définitions, mais pas encore le définiendum des définitions. Une autre simplification que nous avons dû introduire concerne les relations entre documents, la granularité de la cible des citations est encore insuffisante (document au lieu de l'élément de discours).

Références

BISHOP A. P. (1999). Document structure and digital libraries : how researchers mobilize information in journal articles. *Information Processing and Management*, **35**(3), 255 – 279.

DE RIBAUPIERRE H. & FALQUET G. (2011). New trends for reading scientific documents. In *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing*, BooksOnline '11, p. 19–24, New York, NY, USA : ACM.

DE WAARD A., SHUM S. B., CARUSI A., PARK J., SAMWALD M. & SÁNDOR Á. (2009). Hypotheses, evidence and relationships : The hyper approach for representing scientific knowledge claims. In *Proceedings 8th International Semantic Web Conference, Workshop on Semantic Web Applications in Scientific Discourse. Lecture Notes in Computer Science, Springer Verlag : Berlin*.

DJIOUA B. & DESCLES J. (2007). *Indexing documents by discourse and semantic contents from automatic annotations of texts*.

GROZA T., MULLER K., HANDSCHUH S., TRIF D. & DECKER S. (2007). Salt : Weaving the claim web. In *Proceedings of the Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007), Busan, South Korea (Berlin, Heidelberg*.

HANNAY T. (2010). What can the web do for science ? *Computer*, **43**(11), 84–87.

HARMSZE F. (2000). *A modular structure for scientific articles in an electronic environment*. PhD thesis.

IBEKWE-SANJUAN F., SILVIA F., ERIC S. & ERIC C. (2008). Annotation of Scientific Summaries for Information Retrieval. In O. A. . H. ZARAGOZA, Ed., *ECIR'08 Workshop on : Exploiting Semantic Annotations for Information Retrieval*, p. 70–83, Glasgow, Royaume-Uni.

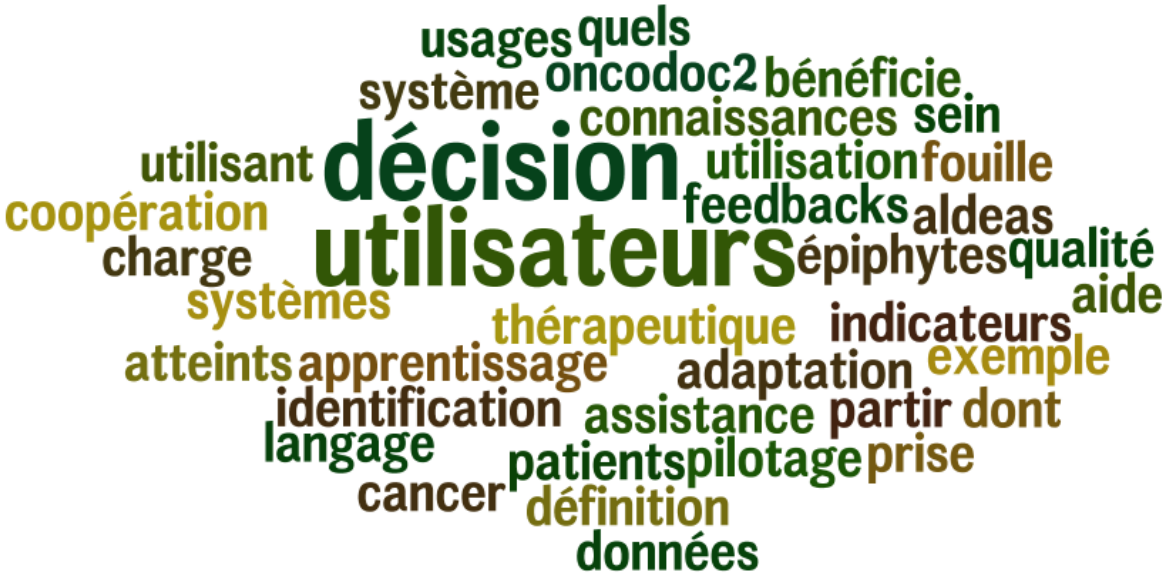
NOVÁCEK V., GROZA T., HANDSCHUH S. & DECKER S. (2010). Coraal - dive into publications, bathe in the knowledge. *J. Web Sem.*, **8**(2-3), 176–181.

RENEAR A. H. & PALMER C. L. (2009). Strategic reading, ontologies, and the future of scientific publishing (vol 325, pg 828, 2009). *Science*, **326**(5950), 230–230.

SHOTTON D. (2009). Cito, the citation typing ontology, and its use for annotation of reference lists and visualization of citation networks. *The 12th Annual BioOntologies Meeting*, p. 1–4.

TEUFEL S. & MOENS M. (2002). Summarizing scientific articles : experiments with relevance and rhetorical status. *Computational linguistics 28*, **4**, 409–445.

Utilisateurs et usages



Quels sont les patients atteints d'un cancer du sein dont la décision de prise en charge thérapeutique bénéficie de l'utilisation d'un système d'aide à la décision ? Un exemple utilisant la fouille de données et OncoDoc2

Jacques Bouaud[1,2,3,4], Arnaud Soulet[5], Jean-Philippe Spano[6,7], Jean-Pierre Lefranc[6,8], Isabelle Cojean-Zelek[9], Brigitte Blaszka-Jaulerry[10], Laurent Zelek[11], Axel Durieux[12], Christophe Tournigand[13], Nizar Messai[5], Alexandra Rousseau[14], Brigitte Séroussi[3,2,4,15]

[1] AP-HP, DRCD, Paris
jacques.bouaud@sap.aphp.fr
[2] INSERM, U1142, LIMICS, Paris
[3] Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1142, LIMICS, Paris
[4] Université Paris 13, Sorbonne Paris Cité, LIMICS, (UMR_S 1142), Bobigny
[5] Université François Rabelais Tours, Laboratoire Informatique EA 6300, Tours
[6] Sorbonne Universités, UPMC Univ Paris 06, UFR de Médecine, Paris
[7] AP-HP, Hôpital Pitié-Salpétrière, Service d'Oncologie médicale, Paris
[8] AP-HP, Hôpital Pitié-Salpétrière, Service de Chirurgie et Cancérologie gynécologique et mammaire, Paris
[9] Hôpital des Diaconesses, Pôle oncologie médicale, Paris
[10] CH Lagny Marne la Vallée, Service de Radiothérapie-Oncologie, Lagny
[11] Université Paris 13, Sorbonne Paris Cité, UFR SMBH, Bobigny ; AP-HP, Hôpital Avicenne, Service d'oncologie médicale, Bobigny
[12] Institut de Cancérologie des Peupliers, Paris
[13] AP-HP, Hôpital St-Antoine, Service d'oncologie médicale, Paris
[14] AP-HP, Hôpital St-Antoine, URC-EST, Paris
[15] AP-HP, Hôpital Tenon, Département de santé publique, Paris ; APREC, Paris

Résumé OncoDoc2 est un système d'aide à la décision (SAD) s'appuyant sur des recommandations de pratique clinique (RPC) pour la prise en charge des cancers du sein. Il a été utilisé comme intervention dans un essai randomisé contrôlé dont l'objectif principal était d'évaluer son impact sur la conformité des décisions des réunions de concertation pluridisciplinaire aux RPC. Nous avons utilisé un algorithme de fouille de données pour découvrir les régularités des profils patients, ou « motifs émergents » (ME), associées à la conformité et à la non-conformité des décisions selon que le système OncoDoc2 était ou non utilisé, afin d'évaluer quels profils patients pouvaient bénéficier de l'utilisation du système. Les ME associés à la non-conformité des décisions prises sans le système sont associées à la conformité quand le système est utilisé sauf dans certaines situations cliniques pour lesquelles la force de la recommandation est faible.

Mots-clés : Recommandations de pratique clinique, conformité aux référentiels, évaluation de la qualité des soins, système d'aide à la décision, fouille de données, cancer du sein.

1 Introduction

Les recommandations de pratique clinique (RPC) sont des synthèses médicales élaborées par les sociétés savantes et les agences nationales dans l'objectif d'améliorer la qualité des soins. Elles proposent pour un ensemble de situations cliniques fréquentes ou complexes, les

principes des prises en charge adaptées en accord avec les dernières données de la science. Ainsi, les RPC devraient réduire les variations inappropriées de pratique, couramment rapportées dans la littérature, optimiser les retombées cliniques et ainsi promouvoir l'efficience. Pourtant, quand elles sont diffusées dans leur format de production sous la forme d'un texte narratif, les RPC ne sont que faiblement suivies par les médecins (Grimshaw et al., 2012). Certaines barrières à leur mise en œuvre ont été effectivement rapportées (Cabana et al., 1999 ; Flottorp et al., 2013). En particulier, les médecins peuvent ne pas connaître ces RPC, ne pas être familier avec leur contenu, être en désaccord avec leur contenu, ou ne pas penser que leur mise en œuvre soit le meilleur choix pour leurs patients.

En rappelant la recommandation de prise en charge adaptée à un patient donné, les systèmes d'aide à la décision (SAD) devraient permettre d'améliorer la conformité des décisions médicales aux RPC, et de nombreuses études ont en effet montré un impact positif de l'utilisation des SAD sur la conformité des pratiques aux RPC. Pourtant, les résultats sont contrastés et il existe des études qui rapportent un effet limité, voire aucun effet de l'utilisation de tels outils sur les décisions des médecins (Shojania et al., 2010 ; Jaspers et al., 2011 ; Grimshaw et al., 2012). Rappeler aux médecins-décideurs les recommandations de prise en charge adaptées au patient, au moment de la décision semble n'être « ni nécessaire ni suffisant » pour garantir la conformité des décisions aux RPC (Shiffman et al., 1999). Aussi, de nombreux travaux cherchent à identifier les déterminants du suivi des recommandations.

OncoDoc2 est un SAD permettant la mise en œuvre des RPC pour la prise en charge des patients atteints d'un cancer du sein non métastatique (Séroussi et al., 2001). OncoDoc2 a déjà été utilisé lors des réunions de concertation pluridisciplinaire de sénologie de l'hôpital Tenon (Paris), au cours d'une étude préliminaire non contrôlée suivant un design avant/après (Séroussi et al., 2007). Le taux de conformité des décisions aux RPC élaborées par CancerEst (fédération des hôpitaux de l'Est parisien impliqués dans la prise en charge du cancer) était significativement plus élevé dans la période après où le système était utilisé (augmentation de 79.2% à 93.4%, $P < 0.0001$). Bien que l'accroissement de la conformité ne puisse pas être strictement imputée à l'impact du système OncoDoc2, les participants à la réunion de concertation pluridisciplinaire et son coordonnateur décidèrent à l'issue de cette étude de continuer à utiliser le système en routine pour chaque décision.. Parallèlement, une étude d'impact d'OncoDoc2 était planifiée incluant six hôpitaux selon un schéma d'essai interventionnel randomisé contrôlé (Haute Autorité de Santé, 2007)[1].

La fouille de données, ou *data mining*, est une activité interdisciplinaire faisant partie des méthodes d'extraction de connaissances à partir de données. Le défi de l'extraction de connaissances à partir des données relève de la gestion des bases de données, des statistiques, de la reconnaissance de formes, de l'apprentissage et de l'intelligence artificielle. Le but est de découvrir automatiquement des informations qui puissent être généralisées en nouvelles connaissances à partir de données existantes. Des techniques de fouille des données ont déjà été appliquées à l'analyse de la conformité des pratiques aux recommandations en alternative aux méthodes statistiques classiques, mais de manière très exploratoire (Svatek et al., 2004 ; Razavi et al., 2008). En particulier, la fouille de modèles (Agrawal & Srikant, 1994) vise à extraire des régularités multi-critères, ou *motifs*, qui satisfont certaines contraintes et définissent ainsi leur pertinence. On appelle motifs émergents (ME), les motifs dont la fréquence est significativement différente entre deux ensembles de données, *i.e.* deux classes (Dong & Li, 1999).

Nous avons déjà utilisé la fouille de données pour identifier les profils cliniques associés à la non-conformité des décisions de prise en charge des cancers du sein. Il s'agissait d'étudier l'utilisation en routine d'OncoDoc2 pour l'aide à la décision des réunions de concertation

[1] Cette étude a été financée par l'Assistance Publique-Hôpitaux de Paris, France (#K070603).

pluridisciplinaire (Séroussi et al., 2012). Dans cet article, nous présentons les résultats de l'utilisation de la fouille de données sur les données issues de l'essai randomisé contrôlé pour identifier les ME associés à la non-conformité dans le groupe d'hôpitaux où OncoDoc2 était effectivement utilisé (bras « intervention » de l'étude) et dans le groupe où le système n'était pas utilisé et les décisions prises comme d'habitude, sans intervention particulière, (bras « contrôle » de l'étude) (Séroussi et al., 2013a). L'objectif est ici de caractériser les profils patients qui sont associés à la non-conformité lorsqu'il n'y a pas d'aide à la décision, mais qui ne le sont pas quand OncoDoc2 est utilisé. On identifie ainsi les patients qui bénéficient de l'utilisation du SAD..

La section suivante présente le système OncoDoc2, ainsi que les méthodes utilisées pour collecter et analyser les données de l'étude. Les résultats sont ensuite décrits dans la section 3. La dernière section interprète et discute les résultats obtenus et les limites de la méthode puis conclue ce travail.

2 Matériel et méthode

2.1 Le système OncoDoc2

OncoDoc2 est un SAD développé selon le paradigme documentaire de l'aide à la décision (Bouaud et al., 1999). Il utilise une base de connaissances structurée sous la forme d'un arbre de décision. Les nœuds représentent les critères cliniques décisionnels. Les arcs représentent les modalités de ces critères. Les chemins de l'arbre de décision représentent ainsi des profils cliniques sous la forme d'une séquence de critères instanciés. La base de connaissances peut être automatiquement exploitée à partir de données patients, à condition qu'elles soient codées, importées, par exemple, d'un dossier patient informatisé. Elle peut également être utilisée de façon interactive, l'utilisateur naviguant au sein de l'arbre de décision et pouvant interpréter et contextualiser les données du patient et les recommandations fournies.

Pour utiliser le système, à partir de la racine de l'arbre de décision et à chaque niveau de profondeur, l'utilisateur répond à des questions qui permettent de renseigner les critères décisionnels (histoire de la maladie, examen clinique, résultats anatomopathologiques, etc.). Les questions sont à choix fermés, et il suffit ainsi de cliquer sur la réponse qui convient (pas de saisie clavier). A la fin de la navigation et comme l'illustre la figure 1, le profil, constitué des critères sélectionnés lors de la navigation, est affiché, et les recommandations correspondant à ce profil sont proposées sous la forme de plans de prise en charge alternatifs. La décision prise par les médecins est alors enregistrée. Lorsque cette décision est choisie parmi les propositions du système, elle est de fait conforme aux recommandations, et la valeur de la conformité est « oui ». Si la décision n'est pas dans les propositions du système, la conformité est « non ».

Dans l'état actuel, la base de connaissances contient 83 critères. L'arbre de décision décrit un total de 47 618 profils cliniques différents (chemins), chacun d'eux conduisant à une et jusqu'à 11 recommandations (3 en moyenne). La caractérisation des profils cliniques requière de deux à 27 critères (19 en moyenne).

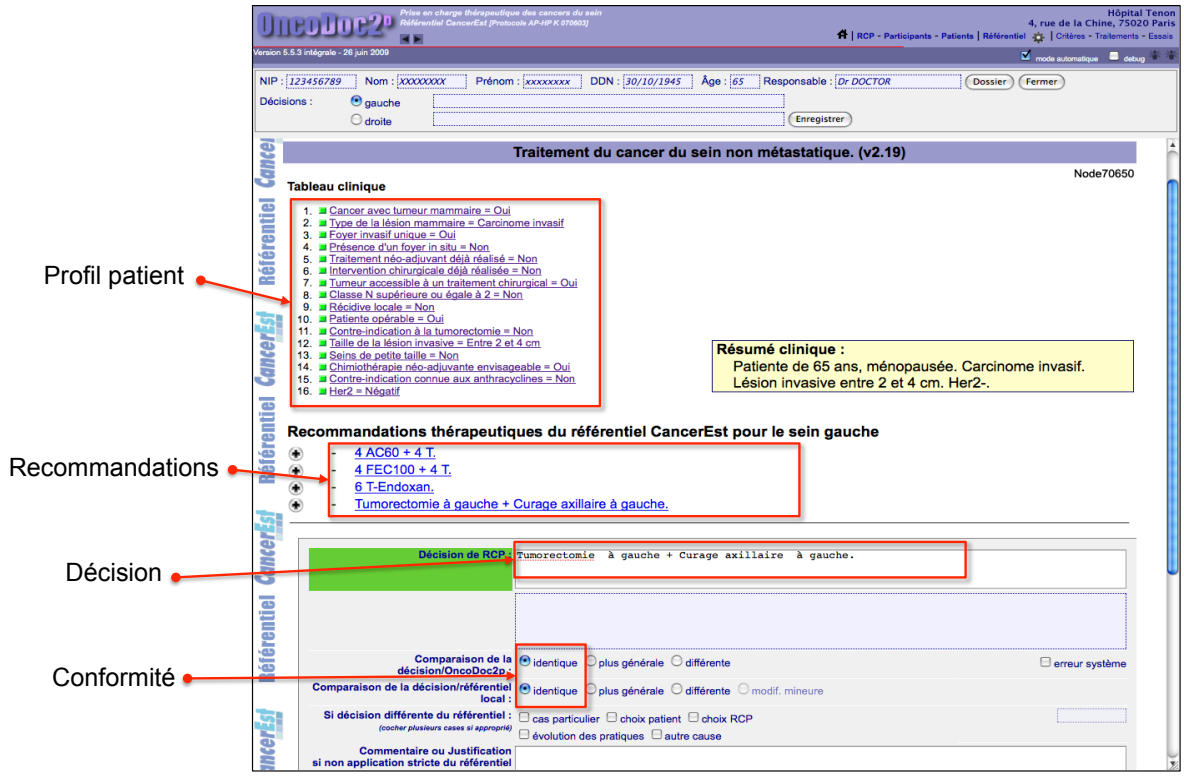


FIGURE 1 – *Copie d'écran de l'interface d'OncoDoc2 lors de l'enregistrement d'une décision.*

2.2 Collecte des données de l'étude

Les six hôpitaux de l'étude sont situés en région parisienne (intra-muros et proche banlieue). Tous organisent des réunions de concertation pluridisciplinaire de sénologie hebdomadaires. Ces réunions associent toutes les spécialités médicales concernées par la prise en charge des patients atteints de cancer (chirurgie, oncologie médicale, radiothérapie, anatomopathologie, radiologie, etc.) afin que les décisions soient prises de façon collégiale et consensuelle sous la forme d'un programme personnalisé de soins optimisé. Pourtant, des études ont montré que ces réunions de concertation pluridisciplinaire avaient à traiter un nombre de cas toujours croissant dans un temps limité. L'attention apporté à chaque cas n'est souvent pas satisfaisante et la conformité des décisions aux RPC n'est pas totalement garantie par ce dispositif organisationnel mis en place par le premier Plan Cancer (Patkar et al., 2012).

Conformément au plan expérimental de l'essai randomisé contrôlé (Haute Autorité de Santé, 2007), le tirage au sort au sein des six structures a permis de déterminer trois hôpitaux qui allaient utiliser OncoDoc2 (bras intervention), et trois hôpitaux qui allaient continuer à fonctionner comme d'habitude, sans SAD (bras contrôle). Dans les hôpitaux avec OncoDoc2, le système a été installé sur l'ordinateur utilisé en réunion de concertation pluridisciplinaire. La vidéo-projection sur grand écran a permis de garantir que chacun des participants des réunions de concertation pluridisciplinaire pouvaient suivre la navigation réalisée en direct au cours de la présentation orale du dossier, et prendre connaissance des différentes recommandations de prise en charge proposées par le système.

Les données recueillies sont celles décrivant le profil clinique des patients : informations relatives au patient (âge au diagnostic, ménopause, traitements déjà administrés, contre-indications à la chirurgie, à la radiothérapie, à la chimiothérapie, etc.), relatives à la tumeur (tumeur invasive associée ou pas à des lésions *in situ*, taille de la tumeur, multifocalité, micro-invasion, récepteurs hormonaux, surexpression HER2, grade SBR, index de prolifération, envahissement ganglionnaire, etc.), et relatives à la chirurgie effectuée (type de la chirurgie mammaire, de la chirurgie axillaire, chirurgie complète ou pas, état des marges, etc.). Les données décrivant la tumeur sont enregistrées au temps pré-chirurgical et au temps post-chirurgical. Les décisions prises par la réunion de concertation pluridisciplinaire ainsi que leur conformité ont été également collectées.

Chaque semaine, des assistants de recherche clinique (ARC) se sont rendus dans chacun des six hôpitaux afin de relever dans les dossiers médicaux les données du profil clinique des patients dont le cas avait été discuté au cours de la réunion pluridisciplinaire de la semaine précédente. Sur la base des dossiers médicaux et pour chaque décision prise, ils procédaient eux-mêmes à une nouvelle navigation OncoDoc2, sans prendre connaissance de celle qui avait été réalisée au cours de la réunion de concertation pluridisciplinaire pour les hôpitaux du bras intervention et relevaient les propositions du système. La conformité de la décision de la réunion de concertation pluridisciplinaire était déclarée avec la valeur « oui » si elle faisait partie des propositions du système récupérées par la navigation des ARC, « non » sinon. Toutes les décisions trouvées non conformes ont été revues dans chacun des six hôpitaux avec l'investigateur local de l'étude. L'objectif était de corriger les éventuelles erreurs de navigation des ARC, et de valider que les décisions étaient effectivement non conformes.

Les données utilisées pour ce travail sont donc celles associées aux décisions des réunions de concertation pluridisciplinaire, c'est-à-dire, pour chaque décision prise, l'ensemble des critères décrivant le profil clinique du patient ainsi que la conformité de la décision aux RPC.

2.3 Principes de la fouille de données

Soit D un ensemble de décisions de réunions de concertation pluridisciplinaire à analyser (Table 1). Chaque ligne de la Table 1 représente une décision caractérisée par des éléments (*e, f, g* et *h)* représentant les valeurs des critères du profil clinique pour lequel la décision a été prise et la conformité. D est divisé en 2 classes, celles des décisions non conformes, D_{non}, et celle des décisions conformes, D_{oui}.

TABLE 1 – *Exemple d'un ensemble D de données correspondant à des décisions.*

N° décision	**Profil**				**Conformité**	**Classe**
d_1	*e*	*f*	*g*		*non*	D_{non}
d_2	*e*	*f*	*g*		*non*	
d_3	*e*		*g*	*h*	*non*	
d_4	*e*	*f*	*g*		*oui*	D_{oui}
d_5	*e*	*f*			*oui*	
d_6		*f*	*g*	*h*	*oui*	

Un motif est un ensemble de critères, par exemple *{e, f, g}* représenté par la suite *efg*. Le motif *X* figure dans la décision d_i si et seulement si *X* est inclus dans d_i. On appelle support du motif *X* dans *D*, et on note *Supp(X;D)*, la fréquence du motif *X* dans *D*, c'est-à-dire, la proportion des décisions de *D* qui contiennent *X*. Par exemple, *Supp(eg;D) = 4/6 et* $Supp(eg;D_{non}) = 3/3$.

La fouille de données par contraste vise à découvrir tous les motifs « intéressants », c'est-à-dire, ceux qui sont présents avec une fréquence significativement plus élevée dans une des deux classes. Le contraste entre deux classes apporté par un motif est mesuré par son taux de croissance ou *growth rate* (GR) (Dong & Li, 1999). Le taux de croissance de D_{non} à D_{oui} d'un motif X est défini par :

$$GR_{non}(X;D) = Supp(X;D_{non}) / Supp(X;D_{oui})$$

Un motif est considéré comme un motif émergeant (ME) si son taux de croissance est supérieur à un certain seuil α. Typiquement, *eg* est un ME avec $\alpha = 2$ parce que son taux de croissance, $Supp(eg;D_{non})/Supp(eg;D_{oui}) = (3/3)/(1/3) = 3$, est supérieur à 2. De la même manière, *eh* est également un ME avec $\alpha = 2$ car son taux de croissance est infini, $Supp(eh;D_{non})/Supp(eh;D_{oui}) = (1/3)/0 = +\infty$. Toutefois, en dépit de son taux de croissance, le motif *eh* n'est pas intéressant car il n'est présent que dans une seule décision. Plus généralement, on considère uniquement les ME « fréquents », c'est-à-dire, les motifs *X* dont la fréquence dépasse un seuil minimal $\gamma > 0$, soit $Supp(X;D) \geq \gamma$.

Il est également possible de filtrer les contrastes qui ne sont pas suffisamment significatifs. Par exemple, avec $\alpha = 2$ et $\gamma = 0.5$, le contraste de *efg* est jugé suffisant car $GR_{non}(efg;D) = 2$ et $Supp(efg;D) = 0.5$. Pourtant, son taux de croissance est faible par comparaison à celui de *eg* ($GR_{non}(eg;D) = 3$). Pour résoudre ce problème, nous calculons une approximation du taux de croissance d'un motif à partir de sa généralisation sur la base d'un ensemble de généralisations formant une partition. Une partition de *X* est un ensemble de motifs $p = \{X_1;\ldots;X_l\}$ tel que les motifs X_i sont deux à deux disjoints, et leur union vaut *X*. L'ensemble de toutes les partitions de *X* est noté *P(X)*. Par exemple, *P(efg)* = *{{e;fg};{ef;g};{eg;f};{e;f;g}}*. Le taux d'inattendu ou *unexpected rate* de *X*, noté UR, mesure la déviation entre le taux de croissance GR et le taux de croissance estimé sur la base d'une hypothèse d'indépendance entre les ensembles de n'importe quelle partition :

$$UR(X;D) = \min_{\{X_1;\ldots;X_n\}\in P(X)} \frac{GR_{non}(X;D)}{\prod \max\{1; GR_{non}(X_i;D)\}}$$

Cette métrique étend le principe introduit par Tati (2010) pour la mesure du taux de croissance. On remarquera que le taux de croissance de chaque généralisation du dénominateur vaut au moins 1 pour ne pas favoriser les motifs dont les généralisations sont faiblement discriminatives. Par exemple, si on considère le motif *eg* dans la Table 1, son taux de croissance estimé est $GR_{non}(e;D) \times GR_{non}(g;D) = 1.5 \times 1.5 = 2.25$, et son taux d'inattendu est $UR(eg;D) = GR_{non}(eg;D)/(GR_{non}(e;D) \times GR_{non}(g;D)) = 3/2.25 = 1.33 > 1$. En d'autres termes, son taux de croissance est supérieur à ce qui était attendu. Un motif est un ME inattendu si son taux d'inattendu est supérieur à un certain seuil ρ. Dans notre exemple, avec $\rho = 1$, *eg* est un ME inattendu.

2.4 Analyse des données

La découverte des « ME d'intérêt » consiste à rassembler les motifs qui vérifient en même temps un support minimal γ, un taux de croissance minimal α, et un taux d'inattendu minimal ρ. Nous avons considéré qu'un motif était « fréquent » si il était présent dans au moins 5 %

des décisions et nous avons posé $\gamma = 5$ %. Nous avons choisi empiriquement d'attribuer la valeur 2 aux seuils α and ρ avec les mêmes paramètres pour permettre la comparaison. Deux ensembles de données ont été construits, D_C pour les données des trois hôpitaux du bras contrôle (sans OncoDoc2), et D_I pour les données des trois hôpitaux du bras intervention (avec OncoDoc2). La fouille de données a été réalisée dans les deux ensembles de données, D_C et D_I. Les ME d'intérêt ont été recherchés à la fois pour leurs sous-ensembles D_{non} et D_{oui}. Nous avons ainsi identifiés les ME associés à la non-conformité (ME-NC) et la conformité (ME-C) dans D_C et D_I.

Les ME qui décrivent les mêmes patients, c'est-à-dire, les motifs qui appartiennent à la même classe d'équivalence, ont été agrégés. Chaque classe d'équivalence est représentée par son motif le plus inattendu, c'est-à-dire le motif qui maximise UR. Un score a été attribué à chaque motif représentant pour quantifier son utilité globale. Parce que nous avons des échantillons de relativement faibles effectifs, nous avons considéré que le support n'était pas un bon indicateur de pertinence (dès lors que la contrainte sur le support en rapport avec le seuil γ était satisfaite). Aussi, pour un motif X, nous avons défini le score d'utilité, noté *score(X;D)*, en considérant le produit $GR(X;D) \times UR(X;D)$. Ainsi, plus le contraste d'un motif est grand et plus il est inattendu, plus son score sera élevé.. Ce score a été utilisé pour ordonner les ME en considérant que plus le score était élevé, plus le ME était intéressant.

3 Résultats

Un total de 825 décisions thérapeutiques a été collecté dans les six hôpitaux de l'étude entre juillet 2009 et avril 2010. Il y avait 268 décisions dans le bras contrôle dont 66 étaient non conformes aux RPC (75 % de conformité). Dans le bras intervention, nous avons collecté 557 décisions dont 65 non conformes (88 % de conformité). Il existe une différence significative entre les taux de conformité mesurés dans les 2 bras, suggérant que l'utilisation d'OncoDoc2 a un impact positif vis-à-vis du suivi des RPC en pratique.

La table 2 présente les résultats obtenus dans chaque ensemble de données D_C et D_I, pour les décisions conformes et non conformes.

TABLE 2 – *Résultats du processus d'identification des ME dans les deux bras de l'étude.*

Jeu de données (décisions)	**Focus**	
	Non conformes	**Conformes**
D_C (sans OncoDoc2)		
# décisions	66	202
# motifs testés	1 623 614	1 623 614
# ME	27	1 384
[score min – score max]	[4,16-54,02]	[4,16-60,28]
D_I (avec OncoDoc2)		
# décisions	65	492
# motifs testés	1 509 049	1 509 049
# ME	4	1 413
[score min – score max]	[4,24-8,29]	[4,04-33,97]

La table 3 présente un extrait des 27 ME-NC de D_C, c'est-à-dire ceux caractérisant la non-conformité des décisions du bras contrôle. La distribution des scores d'utilité associés à ces 27 motifs est représentée dans la figure 2. Nous avons analysé tous les ME situés avant le changement de pente. Le premier ME associé à la non-conformité est « chirinsitub=0 » ce qui correspond à « *chirurgie non carcinologique sur la composante in situ* ». Le second ME correspond aux situations avec « *marges non saines sur la composante in situ* » (marinsitu=1), et les suivants aux situations incluant « *N- et tumeur multifocale* » (nggatt=0, mfocal2b=1).

Toujours dans D_C, le premier ME associé à la conformité sur les 1 384 ME détectés est « proges2=1, mfocal2b=0, meno2=1 » ce qui correspond aux « *patientes ménopausées avec tumeur unifocale et récepteurs à la progestérone positifs* ». Les ME suivants, au sens de leur score d'utilité, décrivent des patientes *« ménopausées avec tumeur unifocale et récepteurs hormonaux positifs* », ou des patientes « *ménopausées ayant eu une chirurgie conservative avec récepteurs hormonaux positif*s », ou des patientes avec « *chirurgie carcinologique sur la composante in situ, sans envahissement ganglionnaire, et sans lésion micro invasive* », ou des patientes « *sans envahissement ganglionnaire, avec récepteurs hormonaux positifs, et HER2-* ». Ce sont pour tous ces cas, des profils de patients bien répertoriés, que l'on peut considérer comme *simples* et pour lesquels la décision thérapeutique est *standard*.

TABLE 3 – *Extrait des 27 ME-NC de D_C, caractérisant la non-conformité dans le bras contrôle.*

N°	Motif	Supp	UR	GR	UR x GR	Longueur
1	chirinsitub=0	0,063	7,35	7,35	54,023	1
	Chirurgie non carcinologique sur la composante in situ					
2	marinsitu=1	0,052	4,08	4,08	16,646	1
	Marges non saines sur la composante in situ					
3	sbr1b=2 nggatt=0 mfocal2b=1	0,052	2,05	7,65	15,683	3
	SBR2 et N- et multifocalité découverte après la chirurgie					
4	er1=1 nggatt=0 mfocal2b=1	0,06	2,27	6,73	15,277	3
	RO+ et N- et multifocalité découverte après la chirurgie					
5	RH1=1 nggatt=0 mfocal2b=1	0,06	2,27	6,73	15,277	3
	RH+ et N- et multifocalité découverte après la chirurgie					
6	hinva1=1 nggatt=0 mfocal2b=1	0,067	2,27	6,12	13,892	3
	Tumeur initiale invasive et N- et multifocalité découverte après la chirurgie					
7	compo1b=1 nggatt=0 mfocal2b=1	0,078	2,21	6,12	13,525	3
	Composante invasive et N- et multifocalité découverte après la chirurgie					
…	…	…	…	…	…	…

Dans D_I, c'est-à-dire au sein des décisions du bras intervention, seuls quatre ME sont associés à la non-conformité. Le premier ME-NC sur les quatre détectés est « cichim=1 » ce qui signifie que « *la chimiothérapie est contre-indiquée* ». Les trois autres ME-NC sont « *exploration de 3 ganglions sentinelles* », « *pas de procédure du ganglion sentinelle et grade SBR1* », ou « *récepteurs aux oestrogènes positifs à 90 %* ».

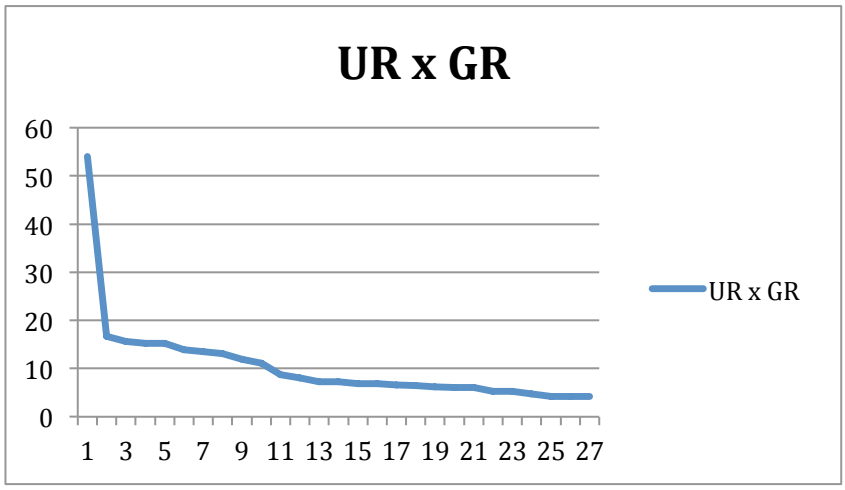


FIGURE 2 – *Courbe du score d'utilité pour les 27 ME-NC de D_C.*

Toujours dans D_I, le premier ME associé à la conformité sur les 1 413 ME détectés est « celtumgs=0 sbr1b=4 RH2=2 » ce qui correspond à des profils patients « *sans cellules tumorales isolées dans les ganglions sentinelles, sans grade SBR pré-chirurgical, et avec récepteurs hormonaux non évaluables* ». Les ME suivants décrivent des profils patients « *sans cellules tumorales isolées dans les ganglions sentinelles* », ou avec *« chirurgie carcinologique sur la composante in situ* », ou des patientes *« ménopausées avec tumeur unifocale et récepteurs hormonaux positifs* », ou encore des patientes avec « *antécédent de mastectomie et récepteurs hormonaux positifs* ».

On notera que les nombres de ME associés à la conformité sont comparables dans D_C (1 384) et D_I (1 413) et relativement élevés, mais que le nombre de ME associés à la non-conformité est très inférieur à la fois dans D_0 et D_1.

4 Discussion et conclusion

Nous avons appliqué une technique de fouille de données afin d'identifier les profils patients associés à la non-conformité des décisions prises en réunions de concertation pluridisciplinaire pour la prise en charge thérapeutique du cancer du sein. Du point de vue de la fouille de données, le nombre de décisions étudiées est faible, en particulier pour les sous-groupes de décisions non conformes. Par ailleurs, les situations où les décisions ne sont pas conformes correspondent à des situations peu fréquentes, hétérogènes et donc plus complexes à caractériser (Liu & Dong, 2012). C'est également probablement la raison pour laquelle elles sont moins facilement reconnues par la méthode développée, en particulier dans le bras intervention, avec OncoDoc2 (D_I). Plus généralement, on peut faire l'hypothèse que certains éléments à l'origine de la non-conformité sont des critères qui ne sont pas pris en compte par les RPC, alors que notre étude est basée sur les critères utilisés dans les RPC.

On trouve un nombre de motifs comparable dans D_C and D_I (1 623 614 et 1 509 049). Il n'est donc pas étonnant d'avoir un nombre quasi identique de ME dans les deux ensembles, et c'est le cas en particulier des ME-C avec seulement 2 % de différence entre D_C et D_I. De façon plus étonnante, la non-conformité est décrite par moins de ME dans D_I que dans D_C. Alors qu'en absolu, il y a quasi autant de décisions non conformes dans les deux ensembles (66 et 65), il y a presque 7 fois plus de ME-NC dans D_C que dans D_I. De plus, les ME de la

non-conformité dans D_I ont des taux de croissance plutôt faibles ce qui indique la pauvreté du contraste. Ce phénomène peut s'expliquer par un effet d'harmonisation des pratiques induit par l'utilisation du système qui élimine les décisions hors-recommandations. En conséquence, les décisions non conformes sont celles pour lesquelles les RPC et donc le système manquent d'éléments décisifs et seraient donc complexes à caractériser.

Nous avons déjà utilisé la fouille de données dans des problématiques analogues à celle présentée dans cet article. Une approche semblable identifiant les ME satisfaisant des contraintes de support, de taux de croissance et de taux d'inattendu a été mise en œuvre sur les décisions prises en utilisant le système OncoDoc2 dans les réunions de concertation pluridisciplinaire d'un seul hôpital (Séroussi et al., 2012). Les résultats étaient différents selon l'étape de traitement. Les patients « *en mauvais état général et avec une contre-indication à la chirurgie* » étaient à risque de décision non conforme dans le groupe pré-chirurgie, les patients avec « *une lésion micro-invasive et une chirurgie non carcinologique* » étaient à risque de décision non conforme dans le groupe reprise chirurgicale, et les patients de *« moins de 35 ans, avec récepteurs hormonaux positifs et HER2+* » étaient à risque de décision non conforme dans le groupe adjuvant. Dans l'étude présentée dans cet article, contrairement à (Séroussi et al., 2012), nous n'avons pas considéré l'étape du traitement, et nous avons rassemblé l'ensemble des données hétérogènes provenant des différentes étapes de la trajectoire thérapeutique, mélangeant ainsi les profils patients. De plus, la fouille de données a été conduite sur les deux bras de l'essai randomisé contrôlé (contrôle et intervention), pour identifier les ME associés à la conformité et à la non-conformité des décisions prises avec, et sans, le système OncoDoc2.

Dans un autre travail précédent, nous avons utilisé l'analyse des concepts formels, une autre technique de fouille de données, pour déterminer les critères patients associés à la conformité et à la non-conformité des décisions prises avec et sans le système OncoDoc2 (Messai et al., 2011 ; Séroussi et al., 2013b). Cette étude avait également été conduite sur un seul hôpital. Elle avait mis en évidence que les décisions prises sans utiliser le système étaient non conformes quand elles concernaient le cas de patients complexes, et que pour certains de ces profils de patients complexes, les décisions devenaient conformes quand les décisions étaient prises avec OncoDoc2. C'est également ce qui est observé dans cette nouvelle étude. Quel que soit le groupe, contrôle ou intervention, les ME associés à la conformité sont des cas simples (p. ex. les patientes ménopausées, avec une tumeur unifocale, des récepteurs hormonaux positifs, sans envahissement ganglionnaire), pour lesquels il n'y a pas de difficulté pour décider de la prise en charge thérapeutique, soit parce que la chimiothérapie n'est pas recommandée ou si elle est recommandée, il s'agit d'un protocole standard. Il en est de même pour la chirurgie, la radiothérapie et les traitements antihormonaux. Pour ces profils patients, il existe des recommandations « fondées sur des preuves », connues et mises en œuvre par les réunions de concertation pluridisciplinaire, qu'elles utilisent ou pas le SAD OncoDoc2. Ainsi, le système n'aurait pas d'impact en terme de conformité sur ce type de profils patients.

Nous avons identifié 27 ME associés à la non-conformité alors que le système n'était pas utilisé (table 3), parmi lesquels figuraient essentiellement les critères patients tels que « *chirurgie incomplète sur la composante in situ* » et « *tumeur multifocale* ». Quand OncoDoc2 était utilisé, les ME associés à la non-conformité étaient différents et beaucoup moins nombreux, puisque seulement quatre ME ont été identifiés. Il s'agit de l'existence d'une « *contre-indication à la chimiothérapie* », des résultats particuliers de la procédure du ganglion sentinelle, et des « *récepteurs aux œstrogènes positifs* ». Les ME associés à la non-conformité dans le bras contrôle n'étaient pas associés à la non-conformité dans le bras intervention. De plus, l'analyse de tous les ME associés à la conformité a permis de mettre en évidence que la prise en charge des tumeurs multifocales était résolue par l'utilisation d'OncoDoc2. En effet, le critère de multifocalité était présent dans cinq ME associés à la conformité quand OncoDoc2 était utilisé (le 303e dans le classement selon le score d'utilité

avec une valeur du score de 12,57, le 472e avec un score d'utilité de 11,29, le 625e avec un score d'utilité de 10,24, le 1015e avec un score d'utilité de 8,53, et le 1120e avec un score d'utilité de 8,12). Ainsi, l'utilisation d'OncoDoc2 a permis d'améliorer la conformité des décisions des réunions de concertation pluridisciplinaire dans le cas des tumeurs multifocales mais pas celles qui concernaient les patients présentant une contre-indication à la chimiothérapie.

Les techniques de fouille de données ont déjà été mises en œuvre pour analyser la conformité des décisions dans la prise en charge de l'hypertension artérielle (Svatek et al., 2004) ou de radiothérapie post-mastectomie (Razavi et al., 2008). Ces deux études visaient à identifier les caractéristiques patients associées à la non-conformité avec différentes techniques de fouilles de données. Ces analyses ont été conduites sur de « vraies » données patients pour des décisions prises sans intervention (c'est-à-dire sans l'utilisation d'un SAD). Dans notre étude, nous avons obtenu des résultats qualitatifs sur les profils patients impactés par l'utilisation du SAD car nous avons pu les comparer dans les deux bras de l'essai randomisé contrôlé quand OncoDoc2 était ou non utilisé. La comparaison permet d'identifier les situations cliniques pour lesquelles les recommandations ont été très faiblement suivies. Ces résultats pourraient être utilisés par les promoteurs des RPC afin de cibler les messages sur les profils à risque de non-conformité et améliorer la qualité des soins. Ils démontrent également que l'utilisation d'un SAD permet de corriger certaines décisions pour ces profils. Néanmoins, il persiste des situations cliniques pour lesquelles les décisions de prise en charge sont non conformes en dépit de l'utilisation d'OncoDoc2. Il s'agit de profils patients non couverts par des recommandations reconnues, pour lesquels il n'existe pas ou peu de preuve ou de consensus et qui doivent être explorés par la recherche clinique.

La poursuite de ce travail consistera à analyser l'impact de l'utilisation d'OncoDoc2 en réunions de concertation pluridisciplinaire par des analyses statistiques classiques et à améliorer la fouille des ME afin de cibler de façon plus fine les critères patients associés à la non-conformité ou à la conformité des décisions. Par ailleurs, les trois seuils utilisés dans la méthode décrite ont été déterminés de façon empirique (comme c'est souvent le cas), en privilégiant les faibles valeurs afin de favoriser la sensibilité de la méthode. L'investigation de techniques permettant d'apprendre ces seuils à partir des données doit être conduite.

Remerciements

Les auteurs remercient les membres des réunions de concertation pluridisciplinaire des centres ayant participé à l'étude, ainsi que l'URCEST pour son support logistique.

Références

AGRAWAL R. & SRIKANT R. (1994). Fast algorithms for mining association rules in large databases. In: Bocca J.B., Jarke M., and Zaniolo C., eds, VLDB. Morgan Kaufmann, p. 487-499.

BOUAUD J., SÉROUSSI B. & ANTOINE É.-C. (1999). OncoDoc : modélisation et "opérationnalisation" d'une expertise thérapeutique au niveau des connaissances. In R. Teulier, editeur, Actes des 3es Journées Ingénierie des Connaissances, p. 61–69.

CABANA M.D., RAND C.S., POWE N.R., WU A.W., WILSON M.H., ABBOUD P.A.C., et al (1999). Why don't physicians follow clinical practice guidelines? A framework for improvement. JAMA ; 282(15), p. 1458-1465.

DONG G. & LI J. (1999). Efficient mining of emerging patterns: Discovering trends and differences. In: KDD ; p. 43-52.

FLOTTORP S., OXMAN A., KRAUSE J., MUSILA N., WENSING M., GODYCKI-CWIRKO M., et al (2013). A checklist for identifying determinants of practice: A systematic review and synthesis of frameworks and taxonomies of factors that prevent or enable improvements in healthcare professional practice. Implement Sci ;8(1), p. 35.

GRIMSHAW J., ECCLES M., LAVIS J., HILL S. & SQUIRES J. (2012). Knowledge translation of research findings. Implement Sci ;7(1), p. 50.

Haute Autorité de Santé (2007). Méthodes quantitatives pour évaluer les interventions visant à améliorer les pratiques. Guide méthodologique, HAS. [Accessible sur www.has-sante.fr].

JASPERS M.W., SMEULERS M., VERMEULEN H. & PEUTE L.W. (2011). Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. J Am Med Inform Assoc ; 18(3), p. 327–334.

LIU Q. & DONG G. CPCQ: Contrast pattern based clustering quality index for categorical data. Pattern Recognition 2012 : 45(4) : 1739-1748.

MESSAI N., BOUAUD J., AUFAURE M.-A., ZELEK L. & SÉROUSSI B. (2011). Using formal concept analysis to discover patterns of non-compliance with clinical practice guidelines: a case study in the management of breast cancer. In: Peleg M., Combi C., Abu-Hanna A., and Andreassen S., eds, AIME, Lecture Notes in Computer Science. Springer, p. 119-128.

PATKAR V., ACOSTA D., DAVIDSON T., JONES A., FOX J. & KESHTGAR M. (2012). Using computerised decision support to improve compliance of cancer multidisciplinary meetings with evidence-based guidance. BMJ Open 2:e000439.

RAZAVI A.R., GILL H., AHLFELDT H. & SHAHSAVAR N. (2008). Non-compliance with a postmastectomy radiotherapy guideline: Decision tree and cause analysis. BMC Med Inform Decis Mak ; 8(1), p. 41.

SÉROUSSI B., BOUAUD J. & ANTOINE É.-C. (2001). OncoDoc, a successful experiment of computer-supported guideline development and implementation in the treatment of breast cancer. Artif Intell Med ; 22(1), p. 43-64.

SÉROUSSI B., BOUAUD J., GLIGOROV J. & UZAN S. (2007). Supporting multidisciplinary staff meetings for guideline-based breast cancer management: a study with OncoDoc2. In: Proc AMIA, Chicago, IL. p. 656-660.

SÉROUSSI B., SOULET A., MESSAI N., LAOUÉNAN C., MENTRÉ F. & BOUAUD J. (2012). Patient clinical profiles associated with physician non-compliance despite the use of a guideline-based decision support system: a case study with OncoDoc2 using data mining techniques. In: Proc AMIA 2012, Chicago, IL. p. 828-837.

SÉROUSSI B., SOULET A., SPANO J.-P., LEFRANC J.-P., COJEAN-ZELEK I., BLASZKA-JAULERRY B., et al. (2013a). Which Patients may benefit from the use of a decision support system to improve compliance of physician decisions with clinical practice guidelines: a case study with breast cancer involving data mining. Stud Health Technol Inform ;192, p. 534-538.

SÉROUSSI B., MESSAI N., LAOUÉNAN C., MENTRÉ F. & BOUAUD J. (2013b). Profils patients associés à la non conformité des décisions aux recommandations de prise en charge thérapeutique des cancers du sein : utilisation de l'analyse de concepts formels. In R. Troncy, éditeur, Actes des 24es Journées Francophones d'Ingénierie des Connaissances, Lille, France, 1–5 juillet 2013.

SHIFFMAN R.N., LIAW Y., BRANDT C.A. & CORB G.J. (1999). Computer-based guideline implementation systems: a systematic review of functionality and effectiveness. JAMIA ; 6(2), p. 104-114.

SHOJANIA K., JENNINGS A., MAYHEW A., RAMSAY C., ECCLES M. & GRIMSHAW J. (2010). Effect of point-of-care computer reminders on physician behaviour: a systematic review. CMAJ ; 182(5), p. 216-225.

SVATEK V., RIHA A., PELESKA J. & RAUCH J. (2004). Analysis of guideline compliance–a data mining approach. Stud Health Technol Inform ; 101, p. 157-161.

TATTI N. (2010). Probably the best itemsets. KDD, p. 293-302.

De la qualité de la coopération à l'identification d'indicateurs de pilotage

Christopher Couthon[1], Régis Martineau[2], Pascal Salembier[1]

[1] ICD/Tech-CICO, UMR 6281 CNRS, Université de Technologie de Troyes
12 rue Marie Curie, BP 2060 - 10010 Troyes Cedex, France,
{christopher.couthon, pascal.salembier}@utt.fr
[2] Ecole Supérieure de Commerce de Troyes,
Campus Brossolette, 217 Avenue Pierre Brossolette, 10000 Troyes
regis.martineau@get-mail.fr

Résumé : De nombreuses organisations utilisent des systèmes de mesures visant à alimenter le suivi et l'évaluation de leurs performances. Ces mesures sont généralement mises en place sous forme d'indicateurs. Ces indicateurs font l'objet de modes de présentation plus ou moins adaptés, sont souvent organisés en tableaux de bord et sont censés donner aux gestionnaires une vision globale pour l'aide à la décision relative à l'anticipation de menaces et à l'optimisation du fonctionnement du système sociotechnique. Dans le contexte du pilotage des systèmes sociotechniques à risques, l'identification d'indicateurs de performance et de suivi du fonctionnement du système est un enjeu majeur. Cette communication présente quelques éléments d'une étude réalisée dans un centre d'appels d'urgence médicale *(SAMU-Centre 15),* préalable au repérage de marqueurs possibles de la « qualité » de la coopération et susceptibles de fournir une information pertinente à une instance de supervision et de pilotage ainsi qu'un retour aux acteurs eux-mêmes sur leur propre activité.

Mots-clés : indicateurs de gestion, qualité de la coopération, mécanismes de coordination.

1 Introduction

De nombreuses organisations utilisent des systèmes de mesure qui visent à alimenter le suivi et l'évaluation de leurs performances. Ces mesures sont généralement mises en place sous forme d'indicateurs préalablement sélectionnés pour leur pertinence au regard du secteur d'activité de l'organisation et de la hiérarchisation de ses priorités (minimisation de la prise de risque, maximisation des résultats productifs, permanence de la qualité de service, etc.). Ces indicateurs font l'objet de modes de présentation plus ou moins adaptés, sont souvent organisés en tableaux de bord et sont censés fournir aux gestionnaires une vision générale et synoptique pour l'aide à la décision relative à l'anticipation de menaces et à l'optimisation du fonctionnement du système sociotechnique.

Les indicateurs de performance utilisés généralement dans les organisations sont orientés résultats et mettent l'accent sur la mesure quantitative des objectifs à atteindre (Kaplan & Norton, 1992). Mais, du fait de leur « incomplétude », ils échoueraient à rendre compte de l'activité réelle des employés (Jordan & Messner, 2012). Malgré cela, les indicateurs de performance ont tendance, dans les organisations où ils sont mis en place, à s'imposer comme critères de décision et d'évaluation. Ce « technicisme » (la croyance en la supériorité de la rationalité technique gestionnaire) constitue une dérive dénoncée par les sociologues de la gestion et les gestionnaires (Boussard, 2008 ; Lorino, 2002), pour ses effets contre-productifs sur le plan organisationnel et néfastes sur le plan humain. C'est pourquoi les organisations ont besoin d'indicateurs davantage orientés vers l'activité, pertinents au regard des tâches effectivement réalisées et des compétences et connaissances mises en œuvre, notamment au niveau de l'activité collective (Engeström, 2000).

Dans le contexte plus spécifique du pilotage des systèmes sociotechniques à risques (incluant les entreprises positionnées sur des secteurs économiques fortement concurrentiels), l'identification d'indicateurs de performance et de suivi du fonctionnement du système est

donc un enjeu majeur. De ce point de vue, l'enrichissement d'indicateurs de gestion classiques (centrés « métier ») par des éléments synthétiques d'évaluation du fonctionnement des collectifs de travail constitue une voie de recherche potentiellement pertinente. L'idée générale est ici d'étudier l'intérêt possible d'intégrer dans des dispositifs d'anticipation et de supervision (de type tableaux de bord notamment) des données de performance classiques et des indicateurs dynamiques fournissant des éléments d'appréciation de l'activité coopérative (distribution de l'information, contexte partagé, alignement des représentations, communications, etc.). Plus fondamentalement l'idée est d'interroger un point de vue traditionnel gestionnaire selon lequel ce type de dispositif pourrait ne reposer que sur un modèle des tâches à réaliser sans référence aux pratiques des acteurs concernés et aux connaissances implicites mises en œuvre.

Cette communication présente quelques premiers éléments d'une étude réalisée dans un centre de traitement des appels d'urgence médicale (SAMU-Centre 15). L'objectif poursuivi est de caractériser les mécanismes de coopération entre acteurs, afin de repérer dans un second temps des marqueurs susceptibles de fournir potentiellement une information pertinente sur la dynamique du fonctionnement du collectif, à une instance de supervision et aux opérateurs eux-mêmes.

2 Problématique

La problématique de la « qualité » de la coopération a été abordée de manière plus ou moins explicite dans différentes communautés de recherche. On citera essentiellement l'apprentissage coopératif assisté par ordinateur (Spada et al., 2005) et l'ergonomie cognitive (par exemple dans les domaines de la conception collaborative et de la gestion des risques : (Burkhardt et al., 2009), (Gaudin et al., 2011). La corrélation entre caractérisation des activités coopératives et niveau de performance métier constitue une question non-triviale (Darcy et al., 2008) : dans certains cas cette relation est quasi-mécanique (cas de certaines activités requérant une forte intégration contrôlée des actions individuelles) ; dans d'autres elle apparaît beaucoup plus ténue à mettre en évidence de manière systématique. Dans ce dernier cas, les connaissances nécessaires à la réalisation du « travail d'articulation », indispensable au déroulement de l'activité coopérative, ne font pas systématiquement l'objet d'un travail de formalisation à visée prescriptive. Un point de vue défendu dans différents champs disciplinaires (l'ergonomie de tradition francophone, les sciences de gestion centrées sur l'activité, le CSCW[1]) pose que l'identification des connaissances, stratégies, compétences qui sous-tendent la mise en œuvre des mécanismes de coordination doit donc passer par l'analyse en situation des pratiques des professionnels, et non pas simplement par l'examen des documents officiels ratifiés par l'organisation (procédures, consignes, « bonnes pratiques »). Toute la difficulté réside ici dans l'interprétation des patterns dynamiques d'activité identifiés : s'agit-il du signe avant-coureur d'un fonctionnement sous-optimal ou « dégradé » du collectif ou s'agit-il d'une réponse de ce collectif à une modification des conditions extérieures (augmentation ponctuelle de la charge de travail, diminution provisoire des ressources humaines disponibles, situation de crise) ?

3 Présentation du terrain de recherche

3.1 Le SAMU 91

Le SAMU (Service d'Aide Médicale Urgente) est un CRRA (Centre de Réception et de Régulation des Appels), intégré au CDAU (Centre Départemental des Appels d'Urgence) du département de l'Essonne (environ 1 200 000 habitants), plate-forme mutualisée avec les Sapeurs-Pompiers (SP) au sein de laquelle sont traités l'ensemble des appels provenant des numéros dédiés 15, 18 et 112. Le principal objectif du SAMU est, à travers la régulation médicale, de déterminer et déclencher la réponse la mieux adaptée à une urgence médicale

[1] Computer-Supported Cooperative Work (Travail coopératif assisté par ordinateur)

dans le délai le plus rapide possible en fonction des ressources (matérielles et humaines) disponibles (SAMU de France, 2009).

Depuis la fin des années 1980, un ensemble de recherches en ingénierie cognitive visant à améliorer l'efficacité du fonctionnement du traitement des appels d'urgence au SAMU 91 ont été menés. Ces travaux ont, entre autres, porté sur la conception d'un collecticiel (Pougès et al., 1994) encore utilisé aujourd'hui dans une version réactualisée, d'un système de dispatching des appels et sur la reconfiguration des espaces de travail suite à la réunion des équipes SAMU et SP qui a abouti à la création du CDAU 91 (Dugdale et al., 2000). L'ensemble de ces travaux se sont basés sur une analyse empirique préalable de l'activité des acteurs et notamment de la dimension coopérative de cette activité (Benchekroun et al., 1995).

3.2 Environnement sociotechnique

Hors situation de crise ou de forte activité, l'équipe en poste au SAMU est composée, en journée, de cinq Assistants de Régulation Médicale (ARM), de deux Médecins Urgentistes Régulateurs[2] (RH), d'un à trois Médecins Généralistes Régulateurs[3] (RL) et d'un ARM dédié aux RH. Ils sont répartis en quatre pôles fonctionnels dans la zone réservée au SAMU 91. Le pôle dit « Mixte » est composé d'ARM installés en face de SP. Le pôle « Santé » est constitué d'un poste d'un RL et d'ARM. Deux autres RL sont regroupés au pôle « Généralistes ». Le pôle « Urgences » comprend deux RH entourant l'ARM dit « dédié ». Le schéma de la figure 1 permet d'illustrer la répartition des tâches de chacun des acteurs du CDAU dans le processus global de régulation des appels d'urgence.

Le poste de travail des ARM et des médecins régulateurs est composé d'un PC avec deux écrans et d'une tablette graphique permettant une prise de notes en temps réel des informations fournies par l'appelant dans un collecticiel, de deux postes téléphoniques (appels entrants et sortants). Une radio connectée au réseau « ANTARES[4] » est disponible seulement sur un poste du pôle « Santé », sur un autre du pôle « Mixte » et sur celui de l'ARM « dédié », mais utilisée prioritairement par ce dernier pour une liaison avec les équipes SMUR[5] en intervention.

4 Approche et méthodologie

Nous avons adopté une démarche en trois phases, La première phase (familiarisation avec le terrain) comprenait deux étapes : une analyse de la tâche prescrite (examen de la documentation disponible et conduite d'entretiens hors situation de travail avec l'encadrement) ; un ensemble d'observations in situ de l'activité de régulation des appels d'urgence au sein de la plate-forme opérationnelle qui nous a permis d'identifier cinq fonctions (RH ; RL ; ARM au pôle « Mixte » ; ARM au pôle « Santé » ; ARM « dédié ») et trois périodes d'activité relativement distinctes (« nominale » ; « pic d'activité pour les RH » ; « pic d'activité pour les RL »).

La deuxième phase (enregistrement audiovisuel de séquences d'activité) a été conduite sur la base des cinq types de fonction et des trois périodes retenues sur chacune des deux équipes d'ARM. Nous avons ainsi effectué 30 sessions d'enregistrement d'une durée de 2 à 3 heures qui ont fait ensuite l'objet d'entretiens d'auto-confrontation avec les agents. Pour ce faire, nous avons mis en place, pour chacune des sessions d'enregistrement, le dispositif multi-sources suivant : deux caméras filmant simultanément l'ensemble du collectif (activité dans la salle de régulation, interactions homme-homme non médiatisées) et un agent en focus (communications verbales et non verbales, interactions médiatisées et non médiatisées localement) selon les fonctions retenues ; un logiciel de capture dynamique d'écran (interactions homme-machine) ; une « double écoute[6] » de l'acteur. L'entrée analytique est ici

[2] Médecins hospitaliers

[3] Médecins de ville libéraux avec un statut de médecin attaché hospitalier

[4] Adaptation Nationale des Transmissions Aux Risques Et aux Secours : cf. http://fr.wikipedia.org/wiki/Antares_(r%C3%A9seau)

[5] Service Mobile d'Urgence et de Réanimation (comprenant un médecin urgentiste, un infirmier et un ambulancier)

[6] Ecoute simultanée des appels téléphoniques reçus par l'acteur en focus lors de la session d'enregistrement

le point de vue d'un acteur dans le traitement coopératif d'un dossier de régulation médicale (DRM).

La troisième et dernière phase sera l'implémentation d'un protocole d'analyse fin de l'ensemble des données collectées, et notamment l'identification de « patterns » de l'activité collective du processus de régulation médicale, en donnant une importance particulière aux mécanismes de coordination. L'objectif sera ensuite de pouvoir associer ces patterns avec des éléments d'appréciation portant sur la « qualité » de la coopération. Dans cet article, nous présenterons uniquement quelques éléments de résultats issus des deux premières phases.

5 Premiers résultats

5.1 Analyse de la tâche prescrite

L'étude de la tâche prescrite de régulation des appels d'urgence met en évidence le modèle de fonctionnement théorique du collectif pensé par l'organisation, au travers de l'allocation des rôles de chacun des acteurs dans la salle de régulation. Le RH doit traiter les demandes dont l'urgence est avérée avec un enjeu vital ou nécessitant des gestes spécialisés et/ou l'envoi d'un SMUR sans perte de temps (suspicions d'arrêt cardio-respiratoire, de détresse respiratoire, de perte de conscience ou coma, d'accident vasculaire cérébral, d'infarctus du myocarde, de traumatismes graves etc.), tandis que le RL doit prendre en charge celles à caractère non vital (conseil thérapeutique comme la posologie d'un médicament, avis médical comme la consultation d'un médecin, la visite d'un médecin, l'envoi d'une ambulance, etc.). L'ARM « dédié » a la responsabilité de déclencher et de coordonner l'envoi des SMUR et des SP. Il doit gérer les dossiers du RH, les suivre en assurant le lien avec les intervenants distants et ainsi décrocher prioritairement les appels provenant des hôpitaux et des SMUR. L'ARM du pôle « Santé » décroche tous les appels, hormis ceux incombant à l'ARM « dédié ». Quant à l'ARM du pôle « Mixte », il doit plutôt répondre aux appels transférés par les SP, sachant que près de la moitié des dossiers de régulation arrivant au SAMU 91 provient des SP (18). Les textes de référence (SAMU de France, 2009), charte interne, fiches de poste, procédures, etc.) ne couvrent explicitement pas ou très peu le travail d'articulation nécessaire et essentiel à la coordination des activités. Les connaissances et compétences que les acteurs doivent mettre en œuvre pour gérer la coopération sur la plateforme sont ainsi méconnues ou ignorées.

5.2 Analyse de l'activité

Le schéma de la figure 1 illustre de manière très simplifiée (de la prise de l'appel au traitement effectif de la demande) le processus de régulation des appels d'urgence tel qu'il a été réellement observé au SAMU de l'Essonne.

L'ARM prend l'appel (étape 0 sur le schéma de la figure 1) et des informations concernant le demandeur (étape 1). Il mène ensuite les premières investigations en dialoguant avec l'appelant (étape 2), afin de qualifier la demande et d'évaluer si un avis médical est requis (étape 3). Dans ce cas, l'ARM transfère l'appel vers (étape 4) un RH pour les urgences vitales (étape 5) ou vers un RL dans le cas contraire (étape 6). S'il n'y a pas besoin d'un avis médical (dans le cas d'une simple demande d'information par exemple), l'ARM traite lui-même directement la demande (étape 7). L'ARM, dit « dédié », contrairement aux autres ARM, ne prend pas en charge les appels en première instance, mais se consacre prioritairement (étape 8) à la mobilisation et à la coordination des divers moyens d'intervention (VSAV[7], SMUR, etc.). Les autres ARM se chargent généralement de l'envoi des ambulances des DRM qu'ils ont initiés (étape 9), spécialement quand l'ARM « dédié » est déjà occupé. D'une manière générale, les ARM alimentent les DRM dans un collecticiel de gestion des appels d'urgence (identité du patient, coordonnées, description de la situation, antécédents et besoins médicaux, etc.). De même, les régulateurs médicaux (RH et RL) y ajoutent leurs décisions médicales.

Le schéma met également en évidence quatre moments de coordination distincts dans l'activité des acteurs, que l'on peut scinder en deux classes : briefing (transmission

[7] Véhicule de Secours et d'Assistance aux Victimes

synthétique des informations importantes et pertinentes issues de la conversation avec l'appelant) du médecin régulateur par l'ARM (notés AA-L et AA-H) ; instructions du médecin régulateur à l'ARM pour l'orientation du patient et l'envoi des moyens de transport le cas échéant (notés AL-A et AH-A).

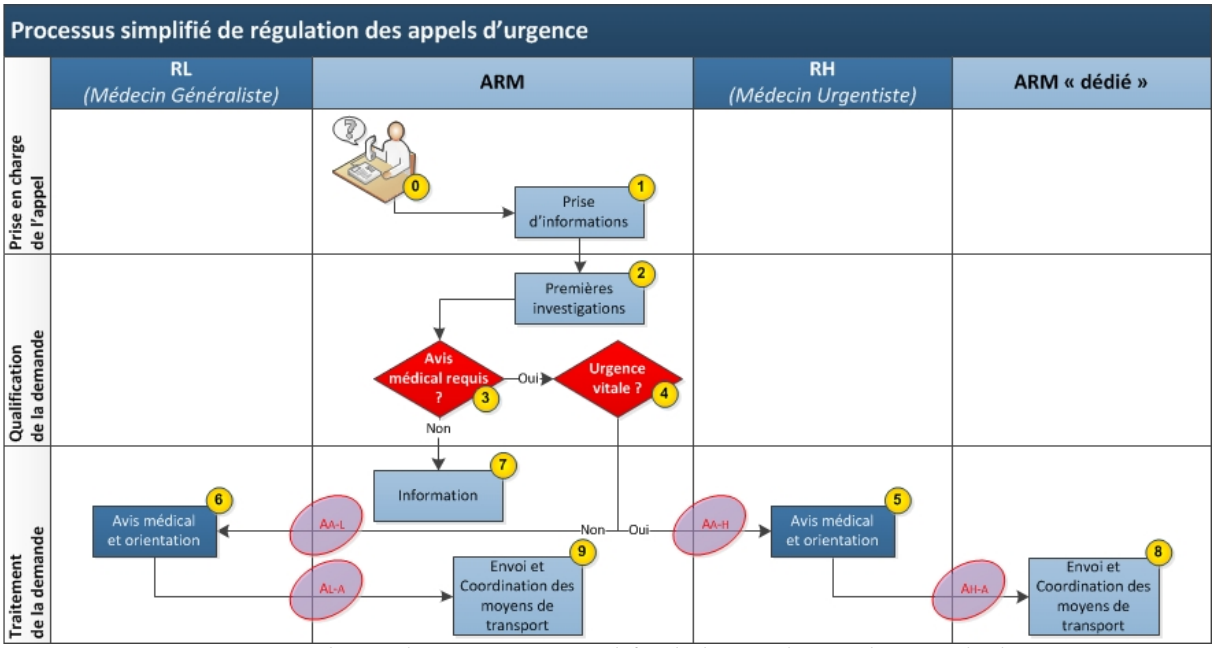


FIGURE 1 – *Schéma du processus simplifié de la régulation des appels d'urgence*

On peut constater que la réalité et les contraintes des situations conduisent, entre autres, à l'émergence d'une utilisation contingente des artefacts et des moyens de communication (vicariance), ainsi qu'à une certaine plasticité dans les rôles des acteurs (allocation/réallocation dynamique de tâches). Nous avons pu relever que même si toutes les conversations doivent être enregistrées via le téléphone (notamment en cas de litige), le degré d'urgence de certaines situations amènent naturellement les acteurs en coprésence à recourir à des communications orales directes non médiatisées. La production d'intelligibilité mutuelle (Salembier & Zouinar, 2004) est ici centrale et repose essentiellement ici sur des mécanismes non intrusifs (écoute flottante) et sur le pluri-adressage des communications. Ceci est particulièrement marqué chez l'ARM « dédié », à l'écoute d'éventuelles instructions des RH, d'alertes données par les autres ARM et des messages radio des équipes d'intervention.

5.3 Ecart constaté

L'écart entre la tâche prescrite et l'activité effective réside essentiellement dans la description de la coopération inter-opérateurs et plus particulièrement des mécanismes de coordination, qui sont généralement fondés sur des connaissances tacites, constituées dans la pratique des acteurs. La formation des ARM repose d'ailleurs non seulement sur des connaissances théoriques paramédicales de base, mais surtout sur une confrontation avec des cas réels sous le tutorat d'un opérateur plus expérimenté.

6 Conclusion, pistes et perspectives

Les premières analyses semblent montrer qu'un pattern global de la coopération émerge de nos observations de terrain (cf. figure 1), représentant la majeure partie des cas traités au SAMU de l'Essonne, le reste ne semblant relever que de cas d'espèce, absorbés la plupart du temps sans difficulté par le collectif. La suite de nos travaux va consister en une analyse plus fine de ces cas pour vérifier si d'autres patterns coopératifs ne permettraient pas d'aller plus avant dans la caractérisation de l'activité. Nous tenterons en outre de vérifier si le recours à ces patterns peut constituer une voie pour la détection de modifications ou d'infléchissements dans le cours de l'activité collective, utilisables comme indicateurs à visée de pilotage. Dans

une optique managériale, ces indicateurs pourront ensuite être regroupés en trois niveaux. Une première batterie d'indicateurs dynamiques synthétiques ou ciblés (Bérard et al., 2009) offrira, au niveau de l'acteur, une vue personnelle, lui permettant de « monitorer » sa propre activité par rapport au collectif en temps réel et, a posteriori, d'utiliser ces indicateurs comme un référentiel. Au niveau du collectif, une vue opérationnelle en situation permettra de doter un « coordinateur[8] » d'indicateurs de suivi à vocation anticipatrice et corrective. Enfin au niveau de l'encadrement, un ensemble d'indicateurs donnera une perspective analytique double : ex ante pour la gestion des ressources et des formations ; ex post pour le retour d'expérience.

Références

BENCHEKROUN, T. H., PAVARD, B., & SALEMBIER, P. (1995). Design of Cooperative Systems in Complex Dynamic Environments. In J.-M. Hoc, C. Cacciabue, & E. Hollnagel (Eds.), *Expertise and technology: cognition & human-computer cooperation* (pp. 167–182). Hillsdale, NJ: LEA.

BERARD, É., GLOANEC, M., & MINVIELLE, É. (2009). Usages des indicateurs de qualité en établissement de santé. *Journal de Gestion et D'économie Médicales*, *27*(1), pp. 5–20.

BOUSSARD, V. (2008). *Sociologie de la gestion : les faiseurs de la performance*. Paris: Perspectives sociologiques, Belin.

BURKHARDT, J.-M., DETIENNE, F., HEBERT, A.-M., & PERRON, L. (2009). Assessing the "Quality of Collaboration" in Technology-Mediated Design Situations with Several Dimensions. In *Proceedings of INTERACT 2009* (pp. 157–160). IFIP International Federation for Information Processing.

DARCY, S., SALEMBIER, P., ANGLEYS, X., BIRAN, H., CARRON, B., & GARDINETTI, E. (2008). Modalités synthétiques d'évaluation / caractérisation des activités coopératives situées - Définition et repérage de marqueurs pertinents. In P. Negroni & Y. Haradji (Eds.), *Actes du congrès de la SELF* (pp. 1–8). Toulouse: Octarès.

DUGDALE, J., PAVARD, B., & SOUBIE, J.-L. (2000). A Pragmatic Development of a Computer Simulation of an Emergency Call Centre. In R. Dieng (Ed.), *Designing Cooperative Systems. Frontiers in Artificial Intelligence and Applications*. IOS Press.

ENGESTRÖM, Y. (2000). Activity theory as a framework for analyzing and redesigning work. *Ergonomics*, *43*(7), pp. 960–974.

GAUDIN, C., DELGOULET, C., GOUNELLE, C., VERNEUIL, L., & BURKHARDT, J.-M. (2011). Evaluation de la qualité de la collaboration lors d'une situation à risque : le cas de la gestion d'un événement NRBC par une équipe multidisciplinaire. In *46e congrès de la Société d'Ergonomie de Langue Française* (pp. 222–228). Issy-les-Moulineaux.

JORDAN, S., & MESSNER, M. (2012). Enabling control and the problem of incomplete performance indicators. *Accounting, Organizations and Society*, *37*(8), pp. 544–564.

LORINO, P. (2002). Vers une théorie pragmatique et sémiotique des outils appliquée aux instruments de gestion. *Essec Research Center*, *DR-02015*.

PENVERNE, Y., JENVRIN, J., DANET, N., PINEAU CARIE, S., POTEL, G., LOUE, B., … BERTHIER, F. (2009). Samu Centre 15 : de nouveaux métiers et nouvelles pratiques. Un centre de réception et de régulation des appels ou de relation patient ? Qualité performance et pilotage. In *3e congrès de la Société Française de Médecine d'Urgence* (pp. 793–806).

POUGES, C., JACQUIAU, G., PAVARD, B., GOURBAULT, F., & CHAMPION, M. (1994). Conception de collecticiels pour l'aide à la prise de décision en situation d'urgence : la nécessité d'une approche pluridisciplinaire et intégrée. In B. Pavard (Ed.), *Systèmes coopératifs : de la modélisation à la conception* (pp. 351–375). Toulouse: Octarès.

SALEMBIER, P., & ZOUINAR, M. (2004). Intelligibilité mutuelle et contexte partagé - Inspirations conceptuelles et réductions technologiques. *@ctivités*, *1*(2), pp. 64–85.

SAMU DE FRANCE. (2009). *Guide d'aide à la régulation au SAMU Centre 15*. Paris: SFEM Editions.

SPADA, H., MEIER, A., RUMMEL, N., & HAUSER, S. (2005). A new method to assess the quality of collaborative process in CSCL. In *Proceedings of the Computer Support for Collaborative Learning Conference* (pp. 622–631). Taïpei, Taïwan.

[8] Nouveau poste en cours de déploiement au SAMU 91, ayant pour vocation de cristalliser la connaissance du collectif et de fluidifier la coopération, appelé « superviseur » dans la littérature médicale (Penverne et al., 2009)

Apprentissage de connaissances d'adaptation à partir des feedbacks des utilisateurs

Abir Beatrice Karami[1], Karim Sehaba[2], Benoît Encelle[1]

[1] Université de Lyon, CNRS
Université Lyon 1
LIRIS, UMR5205, F-69622, France
abir-beatrice.karami@liris.cnrs.fr
benoit.encelle@liris.cnrs.fr
[2] Université de Lyon, CNRS
Université Lyon 2
LIRIS, UMR5205, F-69676, France
karim.sehaba@liris.cnrs.fr

Résumé : Dans le cadre des systèmes adaptatifs, notre travail de recherche porte sur l'acquisition des connaissances d'adaptation à partir des traces d'interaction laissées par les utilisateurs. Les traces contiennent, entre autres, les feedbacks, positifs ou négatifs, des utilisateurs par rapport aux actions du système. Notre objectif est de définir des modèles et des outils permettant d'extraire des règles d'adaptation que le système pourra utiliser, dans son processus de décision, afin de personnaliser son comportement à l'utilisateur. Ces règles d'adaptation établissent des relations de dépendance entre certaines caractéristiques du contexte d'interaction (à savoir certains attributs de la situation d'interaction, tels que le lieu, la luminosité, etc. et/ou du profil de l'utilisateur) et le niveau de satisfaction de l'utilisateur. Pour cela, nous proposons deux algorithmes d'apprentissage. Le premier est direct est certain, dans le sens où toutes les règles générées correspondent à des contextes d'interaction déjà rencontrés par le système, mais nécessite un nombre important de traces pour converger. Le deuxième est plus rapide mais présente des risques d'erreur. En effet, cet algorithme permet de généraliser les règles d'adaptation existantes à de nouveaux contextes d'interaction (i.e. de nouvelles situations et/ou nouveaux profils d'utilisateurs). Dans cet article, nous détaillons les modèles que nous proposons pour représenter les règles d'adaptation et les traces d'interaction ainsi que les deux algorithmes d'apprentissage. Nous présentons également les évaluations que nous avons menées, par simulation et avec de vrais utilisateurs, pour valider nos contributions.

Mots-clés : Systèmes adaptatifs ; traces d'interaction ; apprentissage à partir des feedbacks utilisateurs ; extraction de connaissances d'adaptation.

1 Introduction

L'acquisition des connaissances d'adaptation dans les systèmes adaptatifs est un problème important. En effet, c'est en fonction de ces connaissances que le système s'adapte aux différentes situations auxquelles il est confronté et va faire évoluer ses connaissances. Il s'agit d'un problème complexe et les solutions proposées sont, généralement, dépendantes du domaine d'application.

La plupart des travaux sur l'acquisition des connaissances d'adaptation est basé sur le raisonnement à partir de cas (Wiratunga *et al.* (2002); d'Aquin *et al.* (2007)). Dans (Hanney & Keane (1997)), les auteurs proposent une méthode d'apprentissage des connaissances d'adaptation basée sur la comparaison. D'autres techniques, issues de la génération des recettes de cuisine, ont été développées (Ihle *et al.* (2009); Gaillard *et al.* (2012)). Dans (Gaillard *et al.* (2012)), les connaissances d'adaptation prennent la forme de *substitutions*. Ces dernières sont découvertes à travers un processus de data-mining. Un deuxième processus, piloté par un expert, est nécessaire afin de généraliser ces substitutions et les rendre utilisables dans d'autres recettes. D'autres approches sont basées sur l'analyse des traces d'interaction. Dans (Sehaba (2011)), l'auteur propose une approche permettant à différents utilisateurs de profils très différents, par rapport à

leurs compétences, capacités et/ou préférences, de partager les traces de leurs propres activités. En utilisant un processus de transformation à base de règles, les traces partagées sont adaptées à l'utilisateur cible en fonction de son profil. La génération des connaissances d'adaptation, ici, est basée sur l'application de règles prédéfinies par des experts sur les traces d'interaction.

Une autre approche basée sur les traces d'interaction est proposée (Kanda & Ishiguro (2005)). Cette approche a été appliquée sur un robot compagnon destiné aux enfants. Dans ce cadre, le système développé permet d'identifier chaque enfant et de mémoriser l'historique de ses interactions. En fonction de cet historique, le système est capable de personnaliser ses interactions à l'enfant et aussi de s'adapter à la situation. Néanmoins, les connaissances d'adaptation ici sont prédéfinies et non générées automatiquement. D'autres approches utilisant les feedbacks utilisateurs se sont développées. Dans (Knox & Stone (2009)), les auteurs proposent un framework basé sur des techniques *shaping*. Dans ces techniques, il est nécessaire que l'utilisateur observe le système afin de lui fournir un feedback sur la qualité de ses actions. Ces techniques utilisent des méthodes d'apprentissage supervisé pour générer des fonctions de renforcement. Ces fonctions sont ensuite utilisées par le système, dans son processus de décision, afin de choisir les actions les plus appropriées à l'ensemble des utilisateurs. Même si ces techniques permettent l'apprentissage de certaines tâches, elles ne sont pas adaptées à des environnements interactifs nécessitant des adaptations aux particularités et aux spécificités de chaque utilisateur.

L'objectif général de notre recherche est de développer des modèles et des outils permettant aux systèmes interactifs de s'adapter aux différentes situations et de personnaliser leurs comportements en fonction des préférences et besoins des utilisateurs. Dans ce cadre, nous nous intéressons, en particulier, à l'apprentissage des connaissances d'adaptation à partir des traces d'interaction. Ces dernières contiennent, entre autres, les feedbacks des utilisateurs. Les connaissances d'adaptation ici établissent des relations entre la satisfaction de l'utilisateur et les actions du système, le profil de l'utilisateur et/ou des informations de la situation d'interaction. Le principe de notre approche est de générer, à partir des règles d'adaptation existantes et des traces d'interaction, de nouvelles connaissances applicables à des nouveaux utilisateurs et/ou de nouvelles situations.

Notre approche a été appliquée sur un processus de décision Markovien PDM (Puterman (1994)). Les règles d'adaptation ont été intégrées dans les fonctions de récompense du PDM afin de permettre une planification adaptative. Ce travail entre dans le cadre du projet FUI Robot Populi. Ce projet, financé par le ministère de l'industrie, vise le développement d'un robot compagnon adaptatif (Karami *et al.* (2013)).

Dans la section suivante, nous présentons les concepts généraux de notre approche ainsi que la formalisation que nous proposons. Dans la section 3, nous présentons nos algorithmes d'acquisition de connaissances d'adaptation, utilisées par le système dans son processus de prise de décision. Afin de valider nos contributions, nous avons mené deux expérimentations : la première par simulation et la deuxième avec de réels utilisateurs. La section 4 décrit ces expérimentations ainsi que les résultats obtenus. La dernière section est consacrée à la conclusion et aux perspectives de ce travail.

2 Concepts et formalisations

Nous présentons dans cette section les définitions et formalisations que nous avons introduites pour représenter une situation, un profil utilisateur et les traces d'interaction.

2.1 Attributs et Contexte d'interaction

Nous définissons les *attributs* $\mathcal{AT}$ représentant un *contexte d'interaction* comme l'ensemble des informations caractérisant d'une part le profil de l'utilisateur $\mathcal{AT}^p$ (par exemple : âge, sexe, préférences, habitudes...) et, d'autre part, la situation d'interaction $\mathcal{AT}^s$ (e.g. lieu, luminosité, température...) : $\mathcal{AT} = \mathcal{AT}^p \cup \mathcal{AT}^s$. Chaque attribut $at \in \mathcal{AT}$ possède une valeur appartenant à son domaine de définition (e.g. $sexe \in \{homme, femme\}$). Une décision (qui se concrétise ensuite par une action) peut être liée à un ou plusieurs de ces attributs. Par exemple, la décision *projeter un film d'action* à un utilisateur donné va être essentiellement liée à un attribut "aime_film_d'action" du profil de l'utilisateur, s'il existe. Naturellement, d'autres attributs peuvent influer sur cette décision tels que : l'heure de la journée, le lieu, l'âge de l'utilisateur...

Dans une assistance interactive ou un système compagnon, il est très difficile, voire impossible, de définir manuellement l'ensemble des règles d'adaptation couvrant tous les contextes d'interaction (i.e. tous les profils des utilisateurs potentiels et toutes les situations possibles). C'est pourquoi, nous visons au développement de techniques d'acquisition automatique permettant de déterminer les attributs importants $\mathcal{AT}_{a_i}$ influant sur chaque décision/action $a_i \in \mathcal{A}$ (l'ensemble des décisions/actions). Ces techniques se basent sur les feedbacks (positifs ou négatifs), stockés dans des traces, des utilisateurs par rapport aux expériences passées. Nous notons $\mathcal{AT}_{a_i} \subseteq \mathcal{AT}$ les attributs importants pour a_i.

Une action adaptée doit prendre en compte *le contexte d'interaction* (i.e. la situation courante et les propriétés du profil de l'utilisateur). Les informations du contexte sont représentées par $s \in \mathcal{S}$. Ainsi,

$$\begin{aligned}\mathcal{S} &= \mathcal{AT} = \mathcal{AT}^p \cup \mathcal{AT}^s \\ &= \{(at^{p_1}, \ldots at^{p_n}, at^{s_1}, \ldots at^{s_m}) : at^{p_i} \in \mathcal{AT}^p, at^{s_i} \in \mathcal{AT}^s\}\end{aligned}$$

2.2 Traces d'interaction et règles d'adaptation

Les traces d'interaction et les règles d'adaptation ont des représentations similaires. En effet, leurs deux modèles contiennent un contexte d'interaction $s \in \mathcal{S}$, une action du système $a \in \mathcal{A}$ et une valeur $v \in [-1, +1]$. Formellement, les traces et les règles sont représentées comme suit : $\mathcal{A} \times \mathcal{S} \rightarrow v$. Elles possèdent néanmoins deux différences :

1. Dans une trace d'interaction, v est une valeur numérique représentant le feedback de l'utilisateur (feedback utilisateur suite à une action du système a). Cette valeur est fournie par une fonction prédéfinie $valeur\_feedback$, par exemple : *valeur_feedback (sourire)=1*. En revanche, la valeur de v dans une règle d'adaptation est calculée par un processus d'apprentissage (cf. section 3).
2. Alors que les attributs d'un contexte d'interaction donné s, dans une trace d'interaction, possèdent des valeurs appartenant à leurs domaines respectifs de définition, les valeurs possibles d'attributs, dans une règle d'adaptation, sont décrites à l'aide de contraintes sous forme d'expressions logiques. La figure 1 montre quelques exemples de traces d'interaction et de règles d'adaptation (le symbole * indique que n'importe quelle valeur pour l'attribut est acceptée).

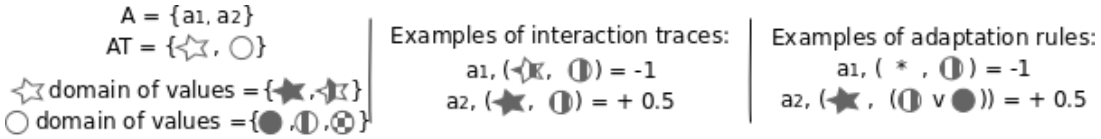


FIGURE 1 – Exemple de traces d'interaction et de règles d'adaptation

Nous précisons qu'une trace d'interaction est *incluse* dans une règle d'adaptation si les valeurs des différents attributs de la trace respectent les contraintes de la règle d'adaptation.

3 Algorithmes d'apprentissage

Dans cette section, nous présentons deux algorithmes d'acquisition de connaissances d'adaptation $\mathcal{K}$ à partir des traces d'interaction. Le premier, direct et certain, permet de générer des règles valables à tout moment. Le deuxième permet de généraliser les règles apprises afin de les appliquer à des nouveaux contextes d'interaction ; c'est-à-dire à des profils et/ou des situations inconnues. Cet algorithme, à la différence du premier, est à risque dans la mesure où il est basé sur la détection automatique des attributs importants $\mathcal{AT}_{a_i}$ pour chaque action a_i. Les deux algorithmes ont principalement comme entrée un ensemble de règles d'adaptation $\mathcal{K}$ apprises à partir des expériences passées (cet ensemble est vide initialement) et une nouvelle expérience, représentée par une nouvelle trace notée $newTrace$, que le système utilise pour mettre à jour ses connaissances d'adaptation $\mathcal{K}$. En sortie, les deux algorithmes retournent un ensemble de règles d'adaptation, éventuellement mis à jour (i.e. ajout de nouvelles règles ou modification de règles existantes dans $\mathcal{K}$).

3.1 Algorithme direct

L'idée principale de l'algorithme 1 est de mettre à jour les règles d'adaptation $\mathcal{K}$ à partir d'une nouvelle trace $newTrace$ de façon simple et directe. Comme le montre cet algorithme, plusieurs situations sont considérées. Les règles d'adaptation concernées par l'apprentissage à partir de $newTrace$ sont celles qui portent sur la même action $a \in A$. Ce qui signifie que pour une nouvelle trace $newTrace : (a_i, s) = value$ seules les règles d'adaptation $\mathcal{K}_a \subseteq \mathcal{K}$ où $a = a_i$ sont considérées. Ainsi :

1. Pour chaque $k \in \mathcal{K}_a$ où $a = a_i$: Si $newTrace$ est *incluse* dans k (Ligne 5), alors :
   (a) Si la valeur de feedback pour $newTrace$ et la valeur de k sont dans la même direction (les deux sont positives ou négatives) (Ligne 6), alors on affecte à la valeur de k la moyenne des deux valeurs (Ligne 7).
   (b) Si les deux valeurs sont de signes différents (une positive et l'autre négative), c'est-à-dire que le feedback de l'utilisateur est en *contradiction* avec une règle existante, $newTrace$ et k seront marquées (Ligne 9 :10). L'ensemble des règles marquées est alors présenté à l'expert afin de détecter les raisons de la contradiction. L'origine de cette dernière peut être le manque d'attributs dans la représentation du problème [1].

1. Par exemple un utilisateur peut changer d'avis si il pleut, d'où l'éventuel besoin d'ajouter un attribut représentant la météo.

2. Si $newTrace$ contient un nouveau contexte d'interaction, n'appartenant à aucune des règles existantes, alors $newTrace$ est ajoutée comme une nouvelle règle d'adaptation dans $\mathcal{K}$ (lignes 11 à 12).

Algorithm 1 Algorithme d'apprentissage direct

```
1: INPUT K, newTrace.
2: Output K.
3: added = false, contradiction_detected = false
4: for all k ∈ K_a do
5:     if (newTrace included in k) then
6:         if (sameDirection(newTrace.value, k.value)) then
7:             k.value = average(newTrace.value, k.value)
8:             added = true
9:         else
10:            contradiction_detected = true
11: if (¬added AND ¬contradiction_detected) then
12:     Add newTrace to K_a
```

3.2 Algorithme de généralisation

L'algorithme 2 a une approche très différente du premier. En effet, il se caractérise par sa capacité à apprendre de nouvelles règles tout en se basant sur un nombre d'expériences moins important. Pour cela, cet algorithme permet de généraliser les règles existantes afin de les appliquer à des nouveaux contextes jusqu'alors inconnus du système.

L'idée principale est de détecter, pour chaque action possible $a \in \mathcal{A}$, l'ensemble des attributs importants (relatifs à l'utilisateur et/ou à la situation) $\mathcal{AT}_a$, c'est à dire ceux dont les valeurs affectent le feedback de l'utilisateur. Les attributs non importants seront alors généralisés à "n'importe quelle valeur " et représentés par des symboles * dans les règles d'adaptation concernées (Ligne 14). Cet algorithme sauvegarde toutes les traces d'interaction. De ce fait, toutes les traces d'interaction sont utilisées, en continu, dans le processus de détection des attributs importants.

Dans les paragraphes suivants, nous expliquons à l'aide d'un exemple, (a) comment l'algorithme utilise les contradictions entre une $newTrace$ et une règle $k \in \mathcal{K}_a$ pour détecter automatiquement les attributs importants et, (b) comment détecter automatiquement les erreurs. Il s'agit des attributs détectés comme étant importants dans la première phase mais, en réalité, non importants.

Initialement, avant la réception d'une première trace $newTrace$, l'ensemble des attributs importants pour chacune des actions possibles est vide ($\mathcal{AT}_a = \emptyset, \forall a \in \mathcal{A}$), et l'algorithme généralise toutes les valeurs des attributs à * dans une nouvelle règle qui correspond à la $newTrace$ (la valeur de la règle produite est alors égale à la valeur du feedback).

Algorithm 2 Algorithme d'apprentissage par généralisation

1: **INPUT** $\mathcal{K}$, $newTrace$, $backupTraces$.
2: **Output** $\mathcal{K}, \mathcal{AT}_a$.
3: **for all** $k \in \mathcal{K}_a$ **do**
4: **if** ($newTrace$ included in k) **then**
5: **if** ($\neg sameDirection(newTrace.value, k.value)$ OR
6: $(sameDirection(newTrace.value, k.value)$ AND
7: $|newTrace.value - k.value| > \epsilon)$) **then**
8: Set $relatedBackup$ as all traces related to k and in the same direction.
9: **for all** $r \in relatedBackup$ **do**
10: **for all** $at \in \mathcal{AT}$ **do**
11: **if** ($k.at = *$ AND $r.at \neq newTrace.at$ AND $at \notin \mathcal{AT}_a$) **then**
12: $\mathcal{AT}_a = \mathcal{AT}_a \cup at$
13: **for all** $k \in \mathcal{K}_a$ **do**
14: Generalize unimportant attributes in k.
15: Specialize important attributes in k from $backupTraces$
16: $k.value = fct(newTrace.value, k.value)$
17: **for all** $at \in \mathcal{AT}_a$ **do**
18: **if** ($\neg$ ConfirmedImportant(at)) **then**
19: $\mathcal{AT}_a = \mathcal{AT}_a - at$.
20: **if** ChangeInImportantAttributes **then**
21: repeat lines

3.2.1 Extraction des attributs importants par traitement des contradictions

Dans les lignes 4 à 7, l 'algorithme 2 examine s'il existe une contradiction entre $newTrace$ et chaque règle $k \in \mathcal{K}_a$. Une contradiction est détectée si $newTrace$ est *incluse* dans k et si leurs valeurs sont opposées (i.e. l'une est positive et l'autre négative) ou sont trop éloignées (la valeur absolue de leur différence est supérieure au seuil ϵ). L'algorithme traite chaque contradiction en vue d'extraire les attributs importants liés à une action a (lignes 8 à 12) comme suit : premièrement, un ensemble non-vide $relatedBackup$ est constitué des traces d'interaction pré-existantes reliées à k et de même direction (i.e. valeurs de même signe). Une trace d'interaction est reliée à k si elle concerne la même action et qu'elle est *incluse* dans k. Ensuite, l'algorithme extrait pour chaque trace de $relatedBackup$ les attributs qui peuvent être source de contradiction avec k (valeurs d'attribut différentes) et les marque comme étant importants. Dans l'exemple donné en Figure 2, après réception d'une deuxième trace, une contradiction est détectée entre cette trace et la règle existante dans $\mathcal{K}_{a_1}$. Cette contradiction est détectée parce que la trace est incluse dans la règle mais a une valeur opposée. Comme indiqué en rouge sur la figure, deux attributs importants sont alors détectés (le cercle et le carré) suite au test défini en ligne 11.

Dans les lignes 13 à 16, l'algorithme commence par généraliser tous les attributs non importants à (*) pour toutes les règles $k \in \mathcal{K}_a$. Ensuite, il spécialise, au regard des attributs nouvellement détectés comme importants, les valeurs des attributs de k en ré-examinant l'ensemble des $backupTraces$. Une règle est alors ajoutée pour chaque trace d'interaction incluse dans k. Au final, la valeur de chaque règle k correspond à la moyenne des valeurs des traces d'interaction

de $backupTraces$ qui sont *incluses* dans k.

3.2.2 Identification des attributs faussement détectés comme importants

L'algorithme peut détecter des attributs comme étant importants alors qu'ils ne le sont pas en réalité. De ce fait, la ligne 18 confirme l'importance réelle de l'attribut.

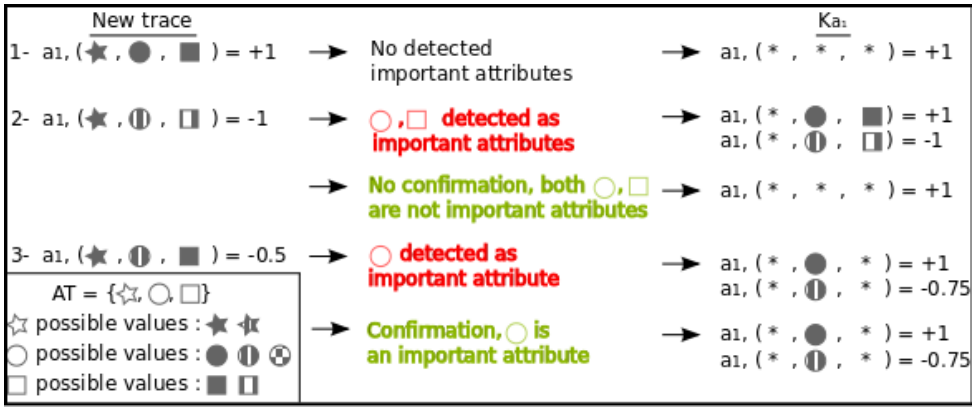


FIGURE 2 – Exemple de détection d'attributs importants : Algorithm 2.

Dans la négative, l'attribut est enlevé de l'ensemble des attributs importants $\mathcal{AT}_a$.

Plus précisément, un attribut important at est confirmé si la condition suivante est vraie pour chaque $k \in \mathcal{K}_a$: il existe au moins une autre règle $k_i \neq k$ où pour tous les autres attributs importants $at_i \neq at$ la valeur de l'attribut dans k_i est égale à sa valeur dans k et les valeurs de ces deux règles sont opposées ou trop éloignées.

Dans l'exemple donné en Figure 2, après réception de la deuxième trace, l'attribut cercle n'est pas confirmé comme étant important (en vert) parce qu'il y a un autre attribut important (le carré) pour lequel il n'y a aucune autre règle ayant la même valeur d'attribut (i.e. le même carré) avec une valeur opposée ou éloignée. De la même manière, l'attribut carré n'est pas confirmé comme étant important. Après réception d'une troisième trace, le cercle est confirmé comme important car la condition est satisfaite (sachant qu'il n'y a pas d'autres attributs importants).

4 Résultats expérimentaux

Dans cette section nous présentons les résultats obtenus après application des deux algorithmes et montrons la capacité de l'algorithme par généralisation à détecter les attributs importants pour chaque action $a \in \mathcal{A}$. Nous avons conduit deux expérimentations. La première a été effectuée par simulation et la deuxième avec des utilisateurs réels.

Ces expérimentations ont été menées dans le cadre du projet FUI Robot Populi. Ce projet vise le développement d'un robot compagnon adaptatif (Karami *et al.* (2013)).

4.1 Expérience en environnement simulé

4.1.1 Scénario

Pour cette expérience, nous considérons une activité de projection de vidéos à l'utilisateur. Cette activité peut être réalisée en suivant les 6 étapes suivantes : sélection 1/d'un pièce pour la projection (chambre principale, chambre secondaire, salon, cuisine), 2/d'une durée (épisode, film), 3/d'un type (dessin animé, manga, science fiction, drame, sport, spectacle, histoire, comédie), 4/d'un volume sonore (faible, moyen, fort), 5/d'une luminosité (faible, moyenne, forte) pour finalement 6/projeter le vidéo. Pour chaque étape, le robot doit décider des actions à réaliser pour que l'utilisateur soit au final satisfait. Dans ce scénario, les attributs représentant l'utilisateur $\mathcal{AT}^p$ sont sa tranche d'âge (enfant, adolescent, adulte) et son sexe (homme, femme). Les attributs modélisant la situation $\mathcal{AT}^s$ sont le niveau de bruit (faible, moyen, élevé), la période de la journée (matin, midi, après-midi, soirée, nuit), la luminosité (faible, moyenne, élevée) et l'étape actuelle dans l'activité.

4.1.2 Procédure

La procédure suivie pour l'évaluation des algorithmes est une boucle, constituée des éléments suivants : (1) Re/calculer la politique de décision du système (politique du MDP). (2) Générer n traces. (3) Evaluer les actions décidées pour chacun des n traces. (4) Extraire et mettre à jour les règles d'adaptation au regard des n traces (mise à jour de la fonction de récompense du MDP). (5) Réitérer ce processus jusqu'à atteindre 2000 traces.

La génération des traces (*i.e.* étape 2) est effectuée par simulation en employant d'une part la politique du MDP pour générer l'action du système et, d'autre part, quelques règles de préférence prédéfinies générant des feedbacks utilisateur. Ces dernières sont prédéfinies par rapport à une action donnée et reliées à certaines valeurs d'attribut du contexte d'interaction (*e.g.* l'action de sélection du type de vidéo dépend de la tranche d'âge et du sexe).

Le comportement du système est généré en utilisant une politique de processus de décision markovien (MDP) calculée en employant un algorithme "value iteration" sur le modèle du MDP (Puterman (1994)). Les états du MDP représentent les contextes d'interaction possibles (i.e. les attributs du profil utilisateur $\mathcal{AT}^p$ et de la situation $\mathcal{AT}^s$). L'ensemble des actions du MDP représente les actions qu'il est possible d'effectuer dans les différentes étapes de l'activité. La fonction de transition change de manière déterministe l'état en fonction de l'action. Une fonction de récompense par défaut est donnée pour respecter l'ordre des étapes permettant de réaliser l'activité (se déplacer dans une pièce, choisir la durée de la vidéo, etc.). La première politique du MDP permet de respecter l'ordre des étapes de l'activité sans inclure des règles d'adaptation.

Dans un premier temps, les traces générées ont à chaque fois respecté les règles de préférence prédéfinies. Il n'y avait pas par conséquent, pour un même contexte d'interaction, de feedbacks utilisateur opposés. Cependant, de telles ambiguïtés (des réactions utilisateur différentes dans un même contexte d'interaction) peuvent exister dans le réel. Par exemple, la plupart mais pas la totalité des adultes hommes aime regarder les émissions sportives. Pour cette raison, dans un second temps, nous avons généré des traces avec une certaine probabilité d'ambiguité (*e.g.* 2% d'ambiguité : une trace est générée avec 2% de chance d'avoir un valeur de feedback utilisateur opposée à celle donnée dans les règles de préférence prédéfinies).

4.1.3 Résultats

Nous avons évalué les deux algorithmes à l'aide de la même base de contextes d'interactions générés (situations et profiles utilisateur générés au hasard). Nous avons suivi la procédure décrite auparavant avec $n = 100$ traces. Après chaque itération (chaque $n = 100$ traces), nous avons calculé le nombre d'actions suivies d'un feedback utilisateur négatif.

Les résultats de la Figure 3 montrent les courbes de convergence des deux algorithmes d'apprentissage vers un comportement adaptatif et personnalisé optimal, i.e. sans actions négatives. Les deux algorithmes ont été capables d'apprendre complètement (à 100%) les règles de préférence prédéfinies. Nous avons mené une expérimentation avec des règles de préférence dépendantes au maximum de deux attributs importants (Figure 3, à gauche), et une autre avec un maximum de trois attributs importants (Figure 3, à droite). Dans les deux cas, les attributs importants prédéfinis ont été appris au cours des expérimentations.

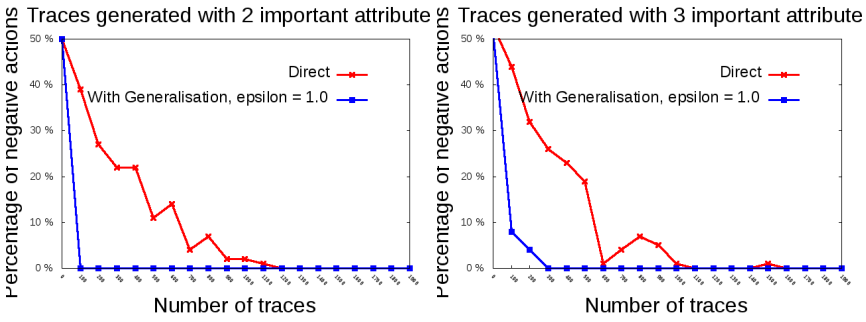


FIGURE 3 – Résultats : traces générées avec deux attributes importants (à gauche) et trois attributs importants (à droite).

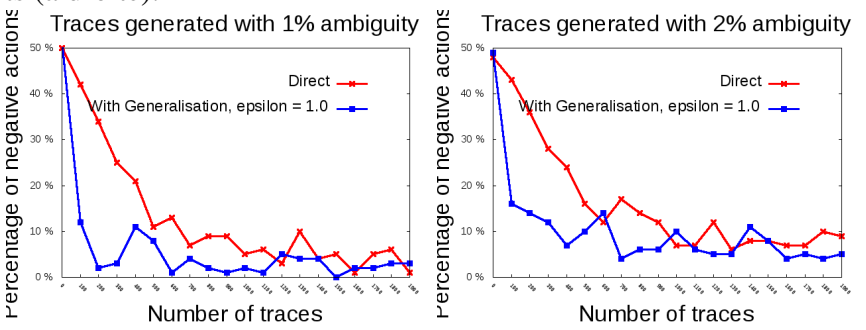


FIGURE 4 – Résultats : traces générées avec 1% d'ambiguité (à gauche) et 2% d'ambiguité (à droite).

Dans la Figure 4, nous exposons les courbes de convergence des deux algorithmes d'apprentissage pour des traces générées à 1% d'ambiguité (à gauche) et à 2% d'ambiguité (à droite). Les résultats montrent que les deux algorithmes n'arrivent pas à converger (avec moins de 2000 traces). Cependant, l'algorithme avec généralisation tend à converger plus rapidement (suite aux 500 premières traces).

4.2 Expérience avec de réels utilisateurs

Dans cette seconde expérience, des utilisateurs réels ont fourni leurs feedbacks suite aux actions du robot. Les traces ont été collectées et employées pour évaluer les algorithmes. Une activité de sélection de menu dans un restaurant à l'aide d'un Robot a été choisie pour cette expérience. Un profil utilisateur est représenté à l'aide de trois attributs : sexe (homme, femme), végétarien (oui/non) et diabétique (oui/non). La situation est représentée à l'aide des attributs suivants : moment du repas (midi, soir) et la saison (hiver ou été).

Nous avons développé une interface permettant à l'utilisateur d'introduire son profil et de préciser la situation. Après avoir donné quelques consignes à l'utilisateur et lorsqu'il est prêt à effectuer sa commande de menu, des questions sont posées de manière séquentielle pour déterminer les choix de l'utilisateur. Ces choix sont ensuite enregistrés dans sa commande afin d'être vérifiés et confirmés.

L'expérience a été réalisée avec 25 sujets adultes (14 hommes et 11 femmes). Face à la difficulté d'avoir des sujets diabétiques et/ou végétariens, chacun des sujets a effectué 4 fois l'expérimentation (i.e. 4 commandes passées au robot). Pour la première commande, le sujet a indiqué son profil réel et pour les trois autres, le système a pour chaque commande indiqué un "faux" profil devant guider ses choix, ceci dans le but de couvrir, pour chaque sujet, l'ensemble des profils possibles (4 cas : végétarien/diabétique, non végétarien/non diabétique, végétarien/non diabétique et non-végétarien/diabétique. Les valeurs des attributs représentant la situation ont été tirées au hasard.

Ces différentes situations ont chacune été distribuées de manière égale en fonction du sexe de l'utilisateur (i.e. autant d'hommes que de femmes ont passé des commandes dans une même situation). Un sujet donné a eu la même situation pour ses quatre commandes.

		M/F	Veg./Non-Veg.	Diab./Non-Diab.	Lunch/Dinner	Summer/Winter
Appetizer + Main Dish	Yes %	52/43	46/50	82/14	52/44	52/44
Appetizer + Dessert	Yes %	0/9	6/2	2/6	0/8	6/2
Main Dish + Dessert	Yes %	11/20	16/14	4/26	15/15	19/12
Appetizer + Main Dish + Dessert	Yes %	38/27	32/34	12/54	33/33	23/42
Green Salad	Yes %	60/63	98/26	60/62	59/64	62/61
Salmon Salad	Yes %	26/23	2/47	27/22	20/30	23/26
Chicken Salad	Yes %	14/14	0/28	12/16	22/7	15/13
Meat Dish	Yes %	30/32	0/61	27/36	33/29	31/31
Vegetables Dish	Yes %	70/68	100/39	73/64	67/71	69/69
Fruits	Yes %	59/68	63/64	100/56	78/52	65/62
Cake	Yes %	41/32	37/36	0/44	22/48	35/38
Red Wine	Yes %	18/9	6/22	4/24	2/25	17/12
White Wine	Yes %	9/5	12/2	2/12	4/10	8/6
Beer	Yes %	7/5	6/6	0/12	8/4	4/8
Only Water	Yes %	66/82	76/70	94/52	85/62	71/75

TABLE 1 – Résultats du premier algorithme (direct) : les réponses des utilisateurs sont catégorisées par valeur d'attribut et exprimées sous forme d'un pourcentage de réponses positives.

Nous présentons dans la Table 1 les pourcentages de réponses positives des utilisateurs (Yes) pour chaque question posée (questions en colonne 1). La table présente ces pourcentages pour les deux valeurs possibles d'un attribut (en colonne, de la colonne 3 à 5). Par exemple, si nous considérons toutes les commandes passées par des hommes pour lesquelles un plat avec viande

Appetizer + Main Dish	diabetes, daytime
Appetizer + Dessert	vegetarianism
Main Dish + Dessert	daytime, diabetes
Appetizer + Main Dish + Dessert	sex, season
Green Salad	vegetarianism, diabetes
Salmon Salad	-
Chicken Salad	diabetes
Meat Dish	vegetarianism, season
Vegetables Dish	-
Fruits	vegetarianism, daytime
Cake	-
Red Wine	sex
White Wine	vegetarianism, diabetes, daytime, season
Beer	diabetes, season
Only Water	-

TABLE 2 – Attributs détectés importants par l'algorithme avec généralisation.

a été proposé (i.e. Meat dish), dans 30% de ces commandes la réponse a été affirmative (le sujet a répondu oui à la proposition d'un plat avec viande).

La Table 2 présente les attributs qui ont été détectés comme importants par l'algorithme avec généralisation pour chaque question posée (i.e. action du robot). Notons ici l'importance des attributs indiquant si l'utilisateur est végétarien ou diabétique, ce qui semble plutôt logique. Par exemple, l'attribut indiquant si l'utilisateur est végétarien ou non a été détecté important pour l'action " Proposer une salade verte" (qui a été choisie par 98% des sujets végétariens, cf. Table 1).

5 Conclusion

Cet article présente des méthodes d'apprentissage qu'un système adaptatif peut employer pour apprendre les préférences de ses utilisateurs à l'aide de leurs feedbacks. Les expériences conduites (par simulation et avec des utilisateurs réels) montrent les capacités de ces méthodes en terme d'extraction de connaissances d'adaptation, afin de personnaliser le comportement d'un système à ses utilisateurs. L'algorithme par généralisation est capable non seulement de traiter des ambiguïtés (feedbacks contradictoires pour un même contexte d'interaction, donnés à des moments différents), mais aussi de déterminer les caractéristiques d'un contexte d'interaction influant sur les feedbacks des utilisateurs.

Comme perspectives, nous aimerions travailler sur la mise à jour des profils utilisateurs en analysant les traces d'interaction (*e.g.* en détectant les préférences personnelles). Nous aimerions également étudier la convergence de l'apprentissage par rapport à la dimension d'un problème (*i.e.* être en capacité de quantifier le nombre de traces nécessaires à un apprentissage de qualité par rapport à la complexité d'un contexte d'interaction donné).

Références

D'AQUIN M., BADRA F., LAFROGNE S., LIEBER J., NAPOLI A. & SZATHMARY L. (2007). Case base mining for adaptation knowledge acquisition. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, p. 750–755, San Francisco, CA, USA.

GAILLARD E., NAUER E., LEFEVRE M. & CORDIER A. (2012). Interactive Cooking Adaptation Knowledge Discovery for the TAAABLE Case-Based Reasoning System. In *Workshop Cooking with Computer - Conférence ECAI 2012*.

HANNEY K. & KEANE M. (1997). The Adaptation Knowledge Bottleneck : How to Ease it by Learning from Cases. In *Proceedings of the International Conference on Case-Based Reasoning.*

IHLE N., NEWO R., HANFT A., BACH K. & REICHLE M. (2009). Cookiis - A Case-Based Recipe Advisor. In S. J. DELANY, Ed., *Workshop Proceedings of the 8th International Conference on Case-Based Reasoning*, p. 269–278, Seattle, WA, USA.

KANDA T. & ISHIGURO H. (2005). Communication robots for elementary schools. *Proceedings of AISB'05 Symposium Robot Companions : Hard Problems and Open Challenges in Robot-Human Interaction (Hatfield Hertfordshire)*, p. 54–63.

KARAMI A. B., SEHABA K. & ENCELLE B. (2013). Towards Adaptive Robots based on Interaction Traces : A User Study . In *The 16th International Conference on Advanced Robotics, ICAR 2013.*, p. 1–6.

KNOX W. B. & STONE P. (2009). Interactively shaping agents via human reinforcement : the tamer framework. In *Proceedings of the fifth international conference on Knowledge capture*, K-CAP '09, p. 9–16, New York, NY, USA : ACM.

PUTERMAN M. L. (1994). *Markov Decision Processes : Discrete Stochastic Dynamic Programming*. New York, NY, USA : John Wiley & Sons, Inc., 1st edition.

SEHABA K. (2011). Partage d'expériences entre utilisateurs différents : adaptation des modalités d'interaction. In *IC 2011 - 22èmes Journées Francophones d'Ingénierie des Connaissances*, p. 639–655.

WIRATUNGA N., CRAW S. & ROWE R. (2002). Learning to adapt for case-based design. In S. CRAW & A. D. PREECE, Eds., *ECCBR*, volume 2416 of *Lecture Notes in Computer Science*, p. 421–435 : Springer.

aLDEAS : un langage de définition de systèmes d'assistance épiphytes

Blandine Ginon[1,2], Stéphanie Jean-Daubias[1,3], Pierre-Antoine Champin[1,3] et Marie Lefevre[1,3]

[1] Université de Lyon, CNRS,
[2] INSA-Lyon, LIRIS, UMR5205, F-69621, France
[3] Université Lyon 1, LIRIS, UMR5205, F-69622, France
{prenom.nom}@liris.cnrs.fr

Résumé : Nous proposons un langage qui permet de spécifier des systèmes d'assistance pour une application-cible donnée, sous la forme d'un ensemble de règles. Ce langage est complété par plusieurs patrons d'actions d'assistance. Nous avons mis en œuvre ces propositions à travers un éditeur d'assistance à destination des concepteurs d'assistance et un moteur générique d'assistance permettant d'exécuter l'assistance spécifiée pour les utilisateurs finaux de l'application-cible sans avoir à la modifier.

Mots-clés : assistance à l'utilisateur, systèmes épiphytes, langage.

1 Introduction

L'assistance aux utilisateurs est l'une des solutions pour pallier les difficultés de prise en main et d'utilisation des applications informatiques. De tels dispositifs permettent d'éviter la sous-exploitation ou le rejet du logiciel.

Nous définissons l'assistance comme l'ensemble des moyens qui permettent de faciliter la prise en main et l'utilisation d'une application, de manière adaptée à l'utilisateur et au contexte d'utilisation. L'assistance vise à permettre à l'utilisateur d'exploiter pleinement toutes les possibilités d'une application, et elle facilite l'appropriation des connaissances et compétences nécessaires à l'utilisation de cette application. Elle comprend les quatre types d'assistance définis par (Gapenne et al., 2002) : substitution, suppléance, assistance et aide.

Le développement d'un système d'assistance adapté à une application est une tâche complexe et coûteuse, souvent négligée par les concepteurs d'applications informatiques. Une personne autre que le concepteur de l'application peut alors souhaiter adjoindre un système d'assistance à une application qui en est dépourvue, ou qui possède un système d'assistance incomplet. Par exemple, dans le cadre d'une communauté d'utilisateurs, un expert peut souhaiter concevoir un système d'assistance pour faire bénéficier de son expérience des utilisateurs plus novices. Le code source de l'application est la plupart du temps non disponible dans le cas où le concepteur de l'assistance n'est pas le concepteur de l'application-cible ; il n'est alors pas possible d'intégrer directement un système d'assistance dans l'application. De plus, comme dans notre exemple, le concepteur potentiel de l'assistance n'est pas toujours un programmeur. Une alternative à l'approche classique de développement d'un module d'assistance intégré dans une application consiste à adopter une démarche épiphyte pour permettre *a posteriori* la spécification et l'exécution d'un système d'assistance dans une application existante sans avoir à la modifier. Un assistant épiphyte est un assistant capable de réaliser une action dans une application-cible externe, sans perturber son fonctionnement (Paquette, et al., 1996).

Le travail présenté dans cet article se situe dans le contexte du projet AGATE (Approche Générique d'Assistance aux Tâches complexEs), qui vise à proposer des modèles génériques et des outils unifiés pour permettre la mise en place de systèmes d'assistance dans des applications existantes, que nous appelons applications-cibles. Nous avons proposé pour cela un processus d'adjonction d'un système d'assistance à une application-cible en deux phases, cf. Figure 1 et (Ginon, et al., 2013b). La première phase concerne un expert de l'application-

cible, appelé par la suite concepteur de l'assistance. Elle permet au concepteur de spécifier l'assistance qu'il souhaite pour une application-cible, en définissant un ensemble de règles d'assistance. La seconde phase concerne les utilisateurs finaux de l'application-cible. Elle consiste en l'exécution de l'assistance souhaitée par le concepteur. Cette phase a lieu à chaque utilisation de l'application-cible par un utilisateur, et se compose de trois processus. La surveillance de l'application-cible (Ginon, et al., 2013a) permet d'observer en continu et de tracer toutes les interactions entre l'utilisateur et l'interface de l'application-cible. En parallèle, le processus d'identification d'un besoin d'assistance exploite les règles d'assistance définies par le concepteur et déclenche le processus d'élaboration d'une réponse au besoin identifié. La réponse se fait sous la forme d'une action d'assistance, réalisée dans l'application-cible par un assistant épiphyte.

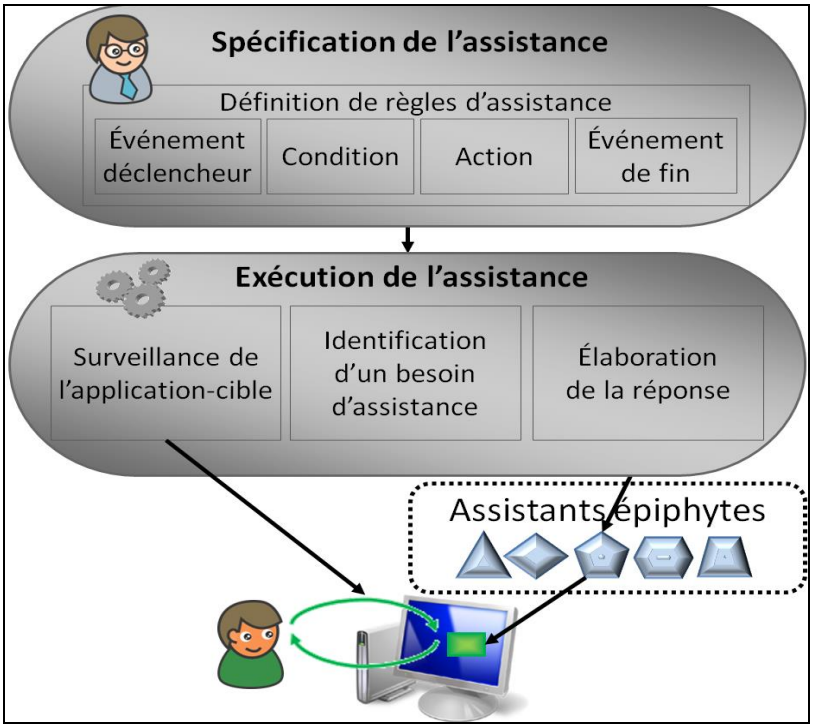


FIGURE 1 – Processus d'adjonction d'un système d'assistance à une application existante.

Dans cet article, nous détaillons aLDEAS, le langage de définition de systèmes d'assistance que nous avons défini afin de permettre la spécification de systèmes d'assistance selon cette approche. Ce langage nous a notamment permis de définir des patrons d'actions d'assistance composées présentés en section 4. Nous présentons ensuite la mise en œuvre de ces patrons dans le système SEPIA. Enfin, nous exposons les évaluations que nous avons effectuées pour ces propositions.

2 État de l'art

La spécification *a posteriori* de systèmes d'assistance pour des applications-cibles existantes a fait l'objet de plusieurs travaux. Les approches de (Paquette, 2012) et (Dufresne, et al., 2003) permettent d'ajouter un système conseiller à un scénario de l'environnement Telos pour le premier et de l'environnement ExploraGraph (Dufresne, 2001) pour le second. Ces systèmes conseillers sont définis par un ensemble de règles de la forme <événement déclencheur, condition de déclenchement, action d'assistance, événement de fin>. Les conditions de déclenchement peuvent inclure une consultation du profil de l'assistance et de

l'historique de l'assistance afin de personnaliser et contextualiser l'assistance. Les actions d'assistance proposées sont de type *message textuel* affiché dans une fenêtre pop-up pour Telos et *animation* ou *message* transmis par un agent animé pour ExploraGraph. L'approche proposée par (Richard, et al., 2004) permet l'ajout d'un système conseiller dans une application Web, afin de permettre le déclenchement d'actions d'assistance lors d'un clic sur un lien. Les actions proposées sont de type *messages textuels* affichés dans une fenêtre popup, les messages peuvent contenir des liens vers une page web ou vers des ressources liées à l'assistance. L'assistance peut être personnalisée en fonction d'un historique de la navigation. Le modèle CAMELEON (Carlier, et al., 2010) permet quant à lui d'ajouter un agent animé capable de se déplacer, de réaliser des *animations* et d'afficher des *messages* dans une application web, grâce à des balises insérées de manière épiphyte dans la page web.

Ces différentes approches ne peuvent cependant pas être utilisées dans n'importe quelle application. Elles sont en effet spécifiques à un environnement donné ou au web. Dans le cadre du projet AGATE, nous nous sommes intéressés à l'adjonction *a posteriori* de systèmes d'assistance dans des applications existantes les plus variées, sans que ces applications soient spécifiques à un domaine ou à un environnement.

3 aLDEAS : un langage de définition de systèmes d'assistance

Pour éviter les confusions avec les styles standards de Word (notamment lors des fusions de documents), les styles spécifiques à la conférence IC2012 sont précédés du préfixe « aic ». En voici la liste et l'usage :

Nous proposons aLDEAS (a Language to Define Epi-Assistance Systems), un langage opérationnalisé dans un outil à destination des concepteurs d'assistance, dont le but est de permettre la définition de systèmes d'assistance sous la forme d'un ensemble de règles, pour des applications existantes. Ce langage est constitué de différents types de composants (cf. Figure 2) que nous présentons de façon détaillée dans cette section. Nous montrerons par la suite comment ces composants peuvent être combinés pour créer des actions d'assistance, qui répondront à des besoins d'assistance.

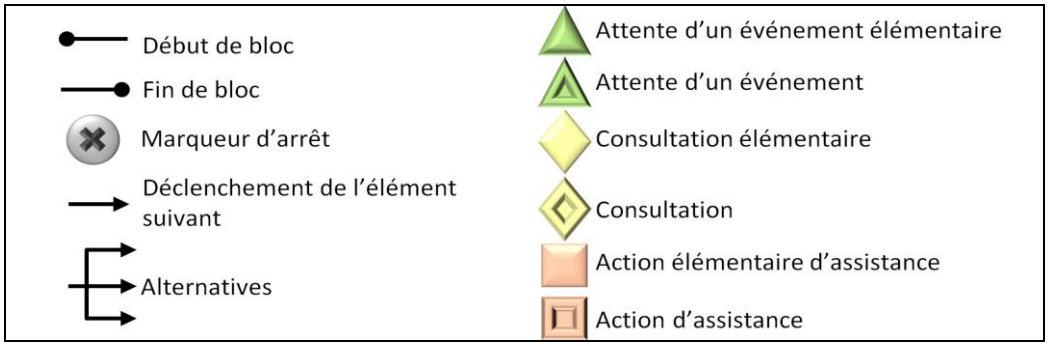


FIGURE 2 – Composants du langage aLDEAS.

3.1 Les attentes d'événements

Notre langage propose des éléments qui permettent d'attendre qu'un événement donné ait eu lieu avant de déclencher un autre élément, telle qu'une action d'assistance. Les attentes d'événements élémentaires, représentées dans le langage par , peuvent être liées à une **action de l'utilisateur**, comme un clic sur un bouton donné, elles peuvent concerner une **action du système d'assistance**, comme le déclenchement ou la fin d'une règle d'assistance, ou la **fin d'un timer**. Un timer est associé à une durée, et peut être déclenché à la suite de n'importe quel événement lié à l'action de l'utilisateur ou à l'assistance. Par exemple, un

timer peut être utilisé pour spécifier qu'une action d'assistance durera 30 secondes. Un timer peut également être utilisé pour définir d'une action d'assistance sera déclenchée si l'utilisateur reste inactif pendant 5 minutes.

Les attentes d'événements élémentaires peuvent être combinées pour former des attentes d'événements composées, représentées par . Ainsi, ces éléments permettent d'attendre une succession donnée d'événements élémentaires formant un événement de plus haut niveau, par exemple une action de correction des yeux rouges sur une photo.

3.2 Les consultations

Le langage aLDEAS propose plusieurs types de consultations élémentaires, représentées dans le langage par , qui peuvent être utilisées pour personnaliser et contextualiser l'assistance. aLDEAS permet la consultation directe de l'utilisateur, pour lui permettre de choisir entre plusieurs options par exemple. il permet également la consultation de l'état de l'application-cible, afin de connaître le texte d'une zone de saisie, ou de savoir quel item est sélectionné dans une liste déroulante par exemple. Enfin, aLDEAS permet la consultation des ressources liées à l'assistance, comme le profil de l'utilisateur qui contient notamment des informations sur les préférences de l'utilisateur en matière d'assistance, l'historique de l'assistance qui contient des informations sur les règles et actions déclenchées pour un utilisateur, et les traces de l'utilisateur qui contiennent des informations sur toutes les interactions entre l'utilisateur et l'interface de l'application-cible.

Les consultations élémentaires peuvent être combinées par une formule logique afin de créer une consultation composée, représentée par . Les consultations, élémentaires ou composées, renvoient une valeur dont le type varie : booléen, texte, ou nombre. Par exemple, une consultation de l'historique de l'assistance peut renvoyer un nombre qui indique combien de fois une action d'assistance a été déclenchée.

3.3 Les actions élémentaires d'assistance

Notre langage propose deux catégories d'actions élémentaires, représentées par : des actions intégrées et des actions extérieures à l'interface de l'application-cible.

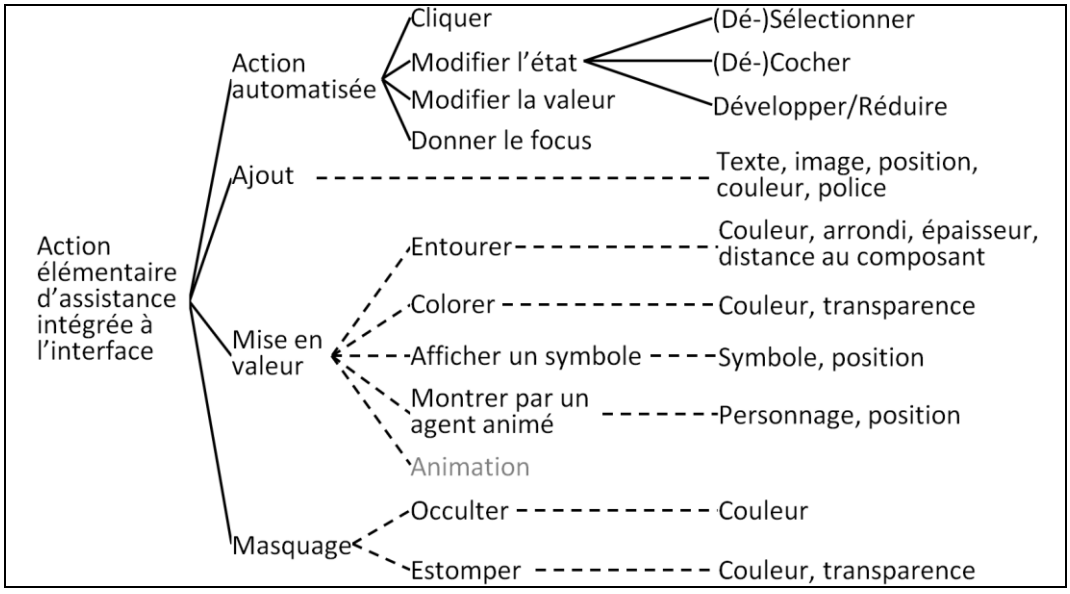


FIGURE 3 – Actions élémentaires d'assistance intégrées à l'interface de l'application-cible.

Les actions intégrées agissent directement sur l'interface de l'application-cible et concernent un composant donné, comme un bouton ou une zone de saisie. Le langage propose quatre types d'actions sur un composant (cf. Figure 3) : **action automatisée**, pour agir à la place de l'utilisateur ; **ajout de composant**, pour enrichir l'interface de l'application-cible, par exemple un bouton permettant de demander de l'aide ; **mise en valeur**, pour guider l'utilisateur et attirer son attention sur un composant ; et **masquage**, pour simplifier aux yeux de l'utilisateur l'interface de l'application-cible. Une action intégrée à l'interface de l'application-cible peut être associée à plusieurs paramètres optionnels, indiqués sur la Figure 3 par des traits en pointillés. Par exemple, pour une mise en valeur, un composant de l'interface de l'application-cible peut être désigné par un agent animé ou entouré par un trait d'une couleur et d'une épaisseur données.

Les actions extérieures à l'interface permettent de proposer à l'utilisateur de l'assistance non associée à un composant de l'interface de l'application-cible. Notre langage en propose trois types (cf. Figure 4) : les **messages langagiers**, associés à un texte pouvant être affiché et/ou lu; les **animations**, par exemple un agent animé qui applaudit; et les **ressources** qui peuvent être proposées à l'utilisateur, par exemple une vidéo de démonstration, un forum ou une application comme la calculatrice.

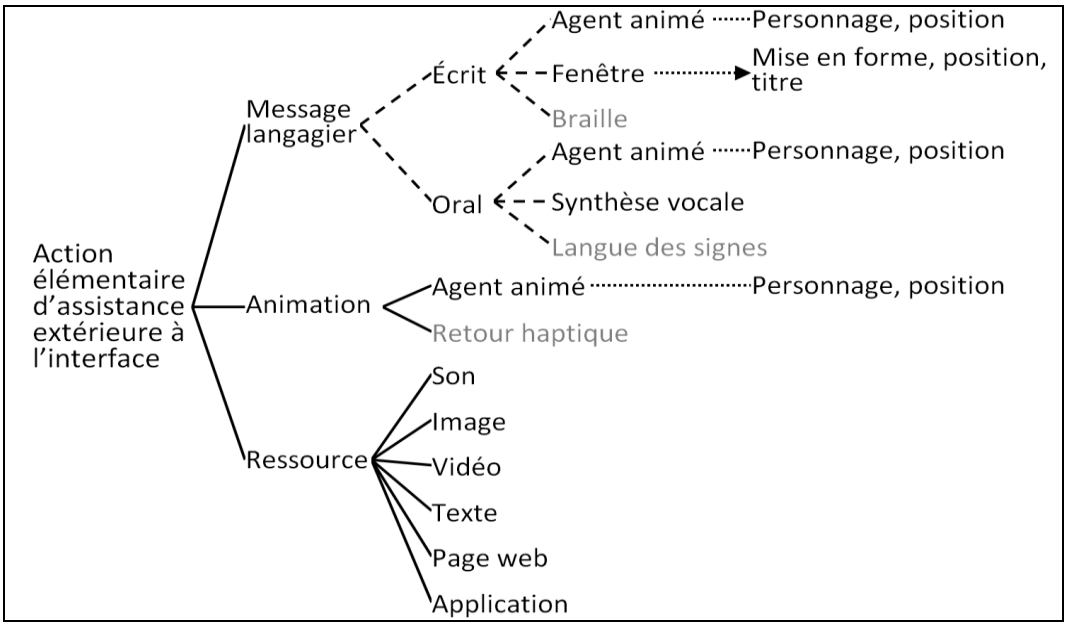


FIGURE 4 – Actions élémentaires d'assistance extérieures à l'interface de l'application-cible.

3.4 Définition d'actions d'assistance composées

Tous les éléments proposés par le langage aLDEAS peuvent être combinés pour créer des actions d'assistance composées, représentées dans le langage par . Par exemple, une action peut être composée d'une séquence d'actions élémentaires ou d'une action associée à un événement de fin, c'est-à-dire une attente d'événement suivie du marqueur qui provoque la fin de toutes les actions lancées depuis le marqueur de début et non encore terminées. La Fig. 5 donne l'exemple d'une action contenant deux actions élémentaires d'assistance : un message et une mise en valeur. Cette action contient également un événement de fin : le message et la mise en valeur disparaîtront au bout de 30 secondes.

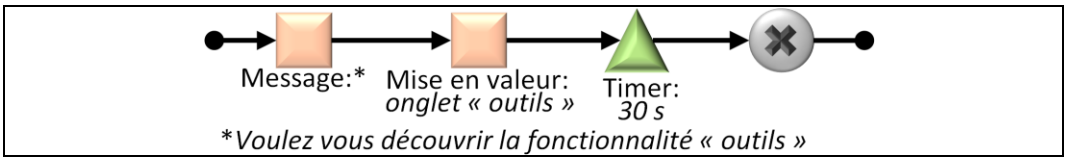


Fig. 5 – Exemple d'une action d'assistance composée.

Dans le cas où une action d'assistance contient plusieurs marqueurs d'arrêt, elle peut renvoyer un état de type texte, indiqué dans le langage sous chaque marqueur d'arrêt (cf. exemple Figure 9 et Figure 10). Les éléments renvoyant une valeur peuvent être suivis d'une alternative à n branches chacune associée à un test dont le type dépend du type de valeur renvoyée. Par exemple pour un texte, les tests peuvent imposer qu'il soit égal, qu'il contienne, commence ou termine par un texte donné. Si un test est vérifié, l'élément suivant de la branche est déclenché. Plusieurs éléments peuvent être déclenchés en parallèle si la valeur de retour vérifie plusieurs tests. Par exemple 4 vérifie $4<7$ et $4 \in [2\ ;\ 9]$. Un exemple d'alternative est donné dans la règle R0 de la Figure 7.

4 Patrons exploitant le langage aLDEAS

Afin de faciliter la définition avec aLDEAS d'actions composées, nous proposons un ensemble de patrons. Pour cela nous enrichissons notre langage avec deux structures supplémentaires : l'embranchement « ou » et les éléments optionnels précédés d'un « ? ». Contrairement aux structures présentées en section 3, celles-ci ne décrivent pas l'exécution de l'assistance ; ce sont des choix à faire par le concepteur de l'assistance au moment de l'instanciation du patron.

4.1 Patron de règles d'assistance

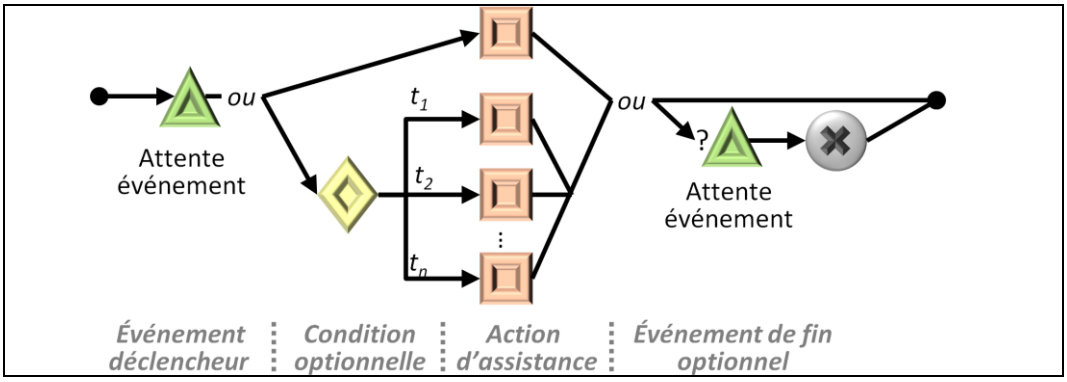


FIGURE 6 – Patron de règles d'assistance.

Le but de notre langage est de permettre à des concepteurs de spécifier par un ensemble de règles l'assistance qu'ils souhaitent pour une application-cible. Dans aLDEAS, nous définissons une règle d'assistance comme une action composée qui respecte le patron donné en Figure 6. Une règle débute par un événement déclencheur qui correspond à l'attente d'un événement. Une règle contient ensuite une action d'assistance ou une consultation avec alternatives associées chacune à une action d'assistance. Une règle peut enfin être associée à un événement de fin, qui correspond à l'attente d'un événement suivi du marqueur d'arrêt. Si une règle n'est pas associée à un événement de fin, elle ne se terminera que lorsque toutes les actions qu'elle a déclenchées se seront achevées. En effet, une action d'assistance peut être associée à son propre événement de fin, et dans certains cas, l'utilisateur peut y mettre fin lui-

même. Par exemple, dans le cas d'un message affiché par une fenêtre pop-up, l'utilisateur peut mettre fin à l'action en supprimant la fenêtre. La Figure 7 donne l'exemple de deux règles d'assistance créées pour l'application-cible PhotoScape1, un logiciel gratuit de retouche d'images. La règle R0 contient un événement déclencheur (le lancement de l'assistance) et une consultation de l'utilisateur lui demandant s'il souhaite de l'aide. R1 sera déclenchée si l'utilisateur choisit l'option « oui, je veux de l'aide ». La règle R7 est déclenchée par le clic de l'utilisateur sur le composant d'identifiant 228 (qui correspond à l'onglet « outils » de PhotoScape). R7 déclenche une action d'assistance constituée de deux actions élémentaires d'assistance : la mise en valeur du composant d'identifiant 210 (le bouton « yeux rouges » de PhotoScape) ; et un message suggérant à l'utilisateur de cliquer sur ce bouton. Le clic de l'utilisateur sur le bouton « yeux rouges » mettra fin à la règle R7 et provoquera donc l'effacement de la mise en valeur et du message.

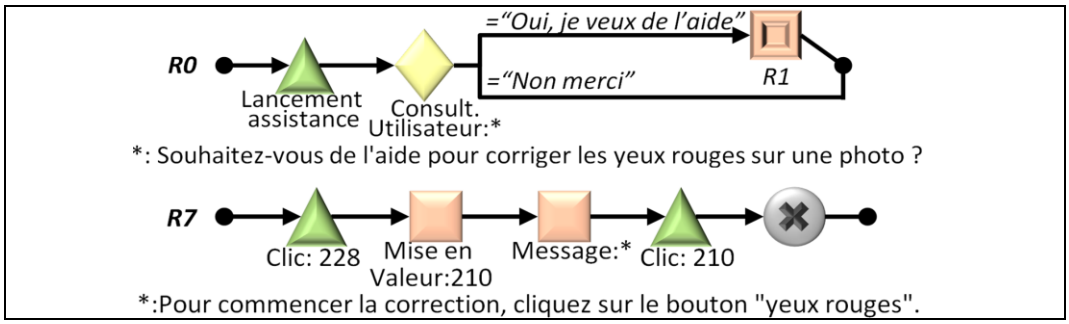


FIGURE 7 – Exemple des règles d'assistance R0 et R7 pour PhotoScape.

4.2 Patron d'actions d'assistance

Le langage aLDEAS permet la définition d'actions d'assistance complexes, combinant de nombreux éléments d'assistance. La définition de telles actions peut être difficile, or aLDEAS s'adresse principalement à des concepteurs d'assistance qui ne sont pas nécessairement informaticiens. Pour cette raison, nous avons défini des patrons d'actions composées, associés à notre langage, afin de faciliter la définition de certaines actions composées fréquemment présentes dans les systèmes d'assistance existants : **actions d'agent animé**, **présentations guidées** et **pas à pas**. Dans cette section, seuls les patrons relatifs aux actions de type pas à pas sont donnés.

Une **action d'agent animé** permet de combiner plusieurs actions élémentaires d'un même personnage : messages, animations (montrer un composant, applaudir, saluer…), et déplacements à l'écran. Par exemple : l'agent animé se place à côté du champ e-mail, il affiche le message « n'oublie-pas de remplir ton adresse mail », il montre le champ e-mail jusqu'à ce que l'utilisateur ait modifié sa valeur.

Une **présentation guidée** comporte plusieurs étapes, dans lesquelles un composant est mis en valeur et éventuellement présenté par un message. On retrouve fréquemment ces actions d'assistance dans les applications existantes, notamment lorsqu'une application est lancée pour la première fois, ou à la suite d'une mise à jour.

Un **pas à pas** vise à faciliter la réalisation d'une tâche en la détaillant sous forme de plusieurs étapes. Chaque étape correspond à une action à réaliser sur un composant de l'interface de l'application-cible. Nous appelons pas à pas automatisé un pas à pas dans lequel le système d'assistance va réaliser les actions à la place de l'utilisateur. Nous appelons pas à pas guidé un pas à pas dans lequel le système va demander à l'utilisateur de réaliser lui-même les actions. Les patrons d'étape de pas à pas automatisé et guidé sont donnés respectivement

[1] http://www.photoscape.org

en Figure 9 et Figure 10 : ces deux patrons d'étapes sont exploités par le patron de pas à pas (cf. Figure 8), sur lequel les étapes sont représentées par .

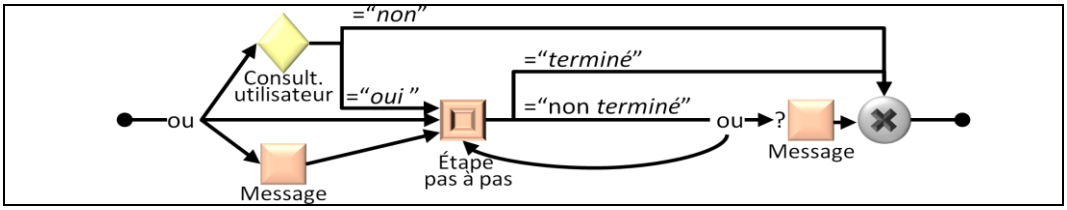


FIGURE 8 – Patron de pas à pas.

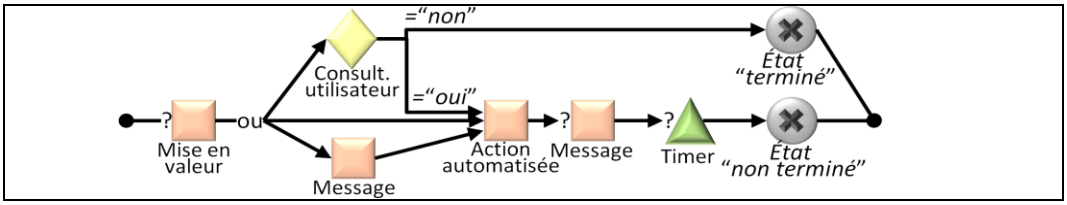


FIGURE 9 – Patron d'étape de pas à pas, en mode automatisé.

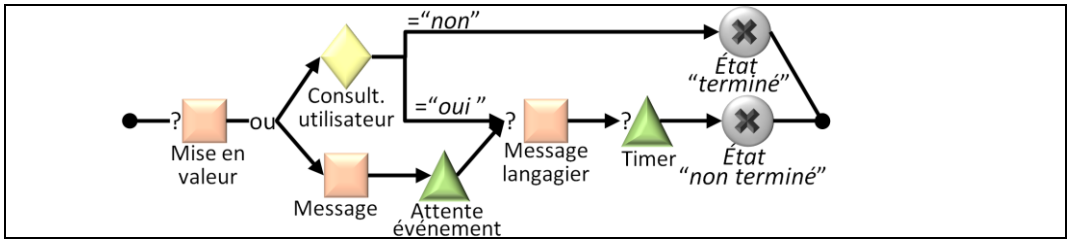


FIGURE 10 – Patron d'étape de pas à pas, en mode guidé.

5 Mise en œuvre du langage dans SEPIA

Nous avons mis en œuvre le langage aLDEAS ainsi que les patrons d'actions d'assistance qui le complètent dans l'environnement SEPIA (Specification and Execution of Personalized Intelligent Assistance).

5.1 L'éditeur d'assistance

L'éditeur d'assistance est un outil destiné aux concepteurs d'assistance. Il met en œuvre aLDEAS et permet de spécifier, pour une application-cible existante, un système d'assistance décrit par un ensemble de règles d'assistance respectant le patron de règles (cf. Figure 6). L'éditeur d'assistance propose une interface pour la création de chaque action élémentaire d'assistance présentée en section 3.3 (à l'exception de celles grisées sur les Figure 3 et Figure 4 qui ne sont pas mises en œuvre dans la version actuelle de SEPIA), ainsi que pour la création d'actions d'assistance qui instancient les patrons que nous avons proposés en section 4.

5.2 Les assistants épiphytes

Nous avons développé un ensemble d'assistants épiphytes, capables de réaliser dans une application-cible les actions élémentaires d'assistance proposées par aLDEAS et définies à

l'aide de notre éditeur. La Figure 11 présente les actions élémentaires d'assistance pouvant être réalisées par nos assistants épiphytes en fonction du type d'application-cible. Pour l'instant, nos assistants épiphytes sont capables d'agir principalement sur les applications Windows natives ou développées en Java, ainsi que sur les applications Web ouvertes avec les navigateurs Chrome ou Firefox.

		Action automatisée	Action mise en valeur	Action masquage	Action message	Action agent animé	Action ressources
Windows	Exécutables	✓	✓	✓	✓	✓	✓
	En Java	✓	✓	✓	✓	✓	✓
	Autres	✗	✗	✗	✓	✓	✓
Web	Firefox	✓	✓	✓	✓	✓	✓
	Chrome	✓	✓	✓	✓	✓	✓
	En Flash	✗	✗	✗	✓	✓	✓
	Autres	✗	✗	✗	✓	✓	✓

FIGURE 11 – Actions élémentaires d'assistance réalisables selon le type d'application-cible.

5.3 Le moteur générique d'assistance

Nous avons développé un moteur générique capable d'exécuter l'assistance spécifiée par le concepteur dans l'éditeur d'assistance. Pour réaliser les actions élémentaires, le moteur fait appel à l'un de nos assistants épiphytes. En ce qui concerne les actions qui instancient l'un des patrons d'actions composées que nous proposons, le moteur assure leur gestion et fait appel à un assistant épiphyte lorsque cela est nécessaire. La Figure 12 montre l'exemple des deux actions élémentaires, une mise en valeur et un message, déclenchées dans l'application-cible PhotoScape par la règle d'assistance R7 (cf. Figure 7 section 4.1).

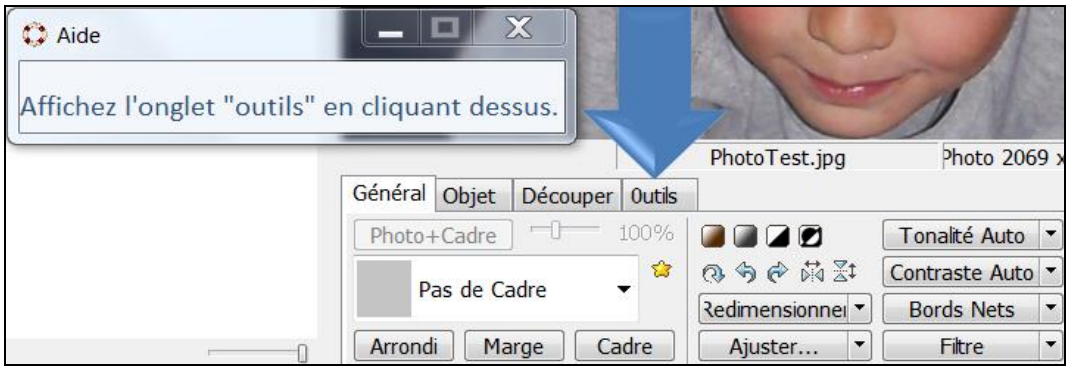


FIGURE 12 – Exemple de deux actions élémentaires d'assistance déclenchées dans PhotoScape.

6 Évaluation et discussion

Les propositions présentées dans cet article ont d'ores et déjà fait l'objet de plusieurs évaluations. En ce qui concerne la faisabilité de notre approche, elle est démontrée par la mise en œuvre que nous en avons faite à travers l'éditeur et le moteur générique d'assistance, complétés par nos collecteurs de traces et nos assistants épiphytes. Dans cette section, nous présentons les études menées pour évaluer la couverture du langage, ainsi que l'utilité et l'acceptation de l'assistance fournie aux utilisateurs finaux.

6.1 Couverture d'aLDEAS

Afin d'évaluer la couverture du langage que nous proposons, nous l'avons utilisé pour modéliser des systèmes d'assistance existants variés, représentatifs des types d'assistance fréquemment rencontrés. Ainsi, notre langage permet de modéliser les systèmes d'assistance souvent utilisés dans les applications Web contenant un formulaire : ils permettent par exemple d'expliquer à l'utilisateur où trouver les informations pour remplir un champ, ou le prévenir qu'un champ n'est pas rempli. Notre langage permet également de modéliser les systèmes d'assistance de type présentation guidée d'un logiciel ou d'une fonctionnalité. Par ailleurs, il existe de nombreux tutoriels proposés par des utilisateurs experts afin de guider un utilisateur pas à pas pour réaliser une tâche, à l'aide de copies d'écran annotées et de messages. Notre langage permet de modéliser de tels systèmes d'assistance avec l'avantage de le faire de façon plus intégrée à l'application-cible. Ainsi, les copies d'écran annotées pourront être remplacées par des actions de mise en valeur intégrée à l'application-cible.

Il existe néanmoins des systèmes d'assistance que notre langage ne permet pas de modéliser : c'est le cas des systèmes d'assistance très spécifiques à une application et requérant des informations non disponibles depuis l'extérieur de cette application. Par exemple, les moteurs de recommandations intégrés dans des applications de vente en ligne utilisent l'ensemble des informations sur les articles consultés ou achetés par tous les utilisateurs du site. Notre langage ne permet pas la consultation de l'ensemble de ces informations.

Nous avons également évalué le langage aLDEAS en le comparant aux approches existantes d'assistance spécifiée *a posteriori* pour une application-cible. En ce qui concerne l'approche proposée par (Richard, et al., 2004), aLDEAS permet également de modéliser de telles actions d'assistance, sous la forme d'une attente d'un événement de type clic sur un lien donné, consultation optionnelle des traces de l'utilisateur relatives à la navigation dans le site, puis action d'assistance de type message (contenant éventuellement des liens vers une autre page web ou ressource). aLDEAS propose également des actions élémentaires d'assistance de lancement de ressources sans passer par un lien. En ce qui concerne l'approche proposée par (Carlier, et al., 2010), aLDEAS permet la définition d'actions d'assistance impliquant des agents animés capables de se déplacer, de s'exprimer oralement et textuellement, ainsi que par des gestes et animations. Nous avons de plus facilité la définition de telles actions en proposant un patron d'action d'agent animé. Enfin, notre langage permet la création de règles de la forme <événement déclencheur, condition de déclenchement, action d'assistance, événement de fin> (cf. section 4.1) équivalentes aux règles utilisées dans les approches de (Paquette, 2012) et (Dufresne, et al., 2003), tout en permettant la définition de règles d'assistance plus complexes et variées. De plus, aLDEAS propose un large choix d'actions élémentaires d'assistance, ainsi que des patrons facilitant la définition d'actions d'assistance composées de nombreux éléments.

6.2 Utilité et acceptation de l'assistance proposée

Afin de tester l'utilité et l'acceptation de l'assistance que nous proposons, nous avons créé à l'aide de SEPIA un système d'assistance pour PhotoScape2. Ce système d'assistance permet de guider les utilisateurs dans la réalisation de chaque étape d'une tâche de correction des yeux rouges. Nous avons demandé à 200 personnes de corriger avec PhotoScape les yeux rouges sur une photo donnée, sans aide pour les 100 utilisateurs du groupe A, et avec l'aide spécifiée pour les 100 utilisateurs du groupe B. Pour le groupe A, on distingue 2 sous-groupes : A1 pour les 49 utilisateurs qui ont réussi à réaliser la tâche demandée sans assistance, et A2 pour les 51 utilisateurs ayant abandonné et à qui nous avons ensuite demandé de réaliser la même tâche avec l'assistance conçue. L'expérimentation était précédée et suivie d'un questionnaire. Les Figure 13 et Figure 14 présentent une partie des résultats de

[2] Vidéo de démonstration est disponible à http://liris.cnrs.fr/blandine.ginon/PhDWork.html

cette étude. On constate notamment que l'aide proposée permet de réaliser la tâche de manière deux fois plus rapide en moyenne. Dans le groupe A1, on note que la moitié des utilisateurs ayant réussi à réaliser la tâche sans assistance aurait pourtant souhaité recevoir de l'assistance. Beaucoup ont en effet précisé sur le questionnaire qu'ils auraient souhaité être guidés pour trouver plus rapidement les composants de PhotoScape permettant la correction des yeux rouges. Dans le groupe A2, les utilisateurs qui n'avaient pas réussi à réaliser la tâche demandée, on constate que 100% d'entre eux ont ensuite réussi grâce à l'assistance proposée et tous ont trouvé l'aide utile. Dans le groupe B, 97% des utilisateurs ont trouvé l'aide utile. L'assistance a été appréciée par les utilisateurs l'ayant testée, on note en effet que 92% des utilisateurs du groupe A2 et 87% des utilisateurs du groupe B l'ont appréciée. Ces résultats très satisfaisants nous conduisent à considérer que le système SEPIA permet de fournir aux utilisateurs finaux d'applications-cibles une assistance à la fois pertinente et efficace pour répondre aux besoins des utilisateurs, tout en étant bien acceptée par ces derniers.

	Effectif	Tâche sans assistance		Tâche avec assistance	
		Taux de réussite	Durée moyenne	Taux de réussite	Durée moyenne
Groupe A	100	49%	-	-	-
A1	49	100%	146,1 s	-	-
A2	51	0%	-	100%	72,15 s
Groupe B	100	-	-	100%	65,33 s

FIGURE 13 – Quelques résultats relatifs à l'expérimentation avec PhotoScape.

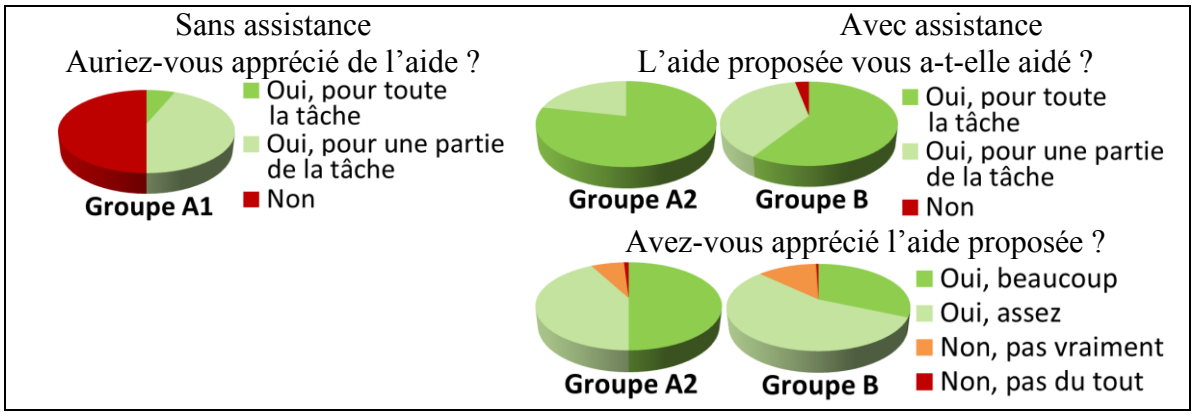


FIGURE 14 – Résultats de questions consécutives à l'expérimentation avec PhotoScape.

7 Conclusion et perspectives

Le langage aLDEAS et les patrons d'actions composées que nous avons présentés permettent la définition de systèmes d'assistance adaptés à leur application-cible. En adoptant une démarche entièrement épiphyte, nous permettons l'adjonction *a posteriori* de systèmes d'assistance à des applications existantes, non conçues spécifiquement pour l'intégration d'assistance, sans accès à son code source et sans connaissances particulières de programmation, par des concepteurs d'assistance qui ne sont pas nécessairement les concepteurs de l'application-cible et qui peuvent être des utilisateurs experts de l'application. Les outils qui mettent en œuvre aLDEAS permettent aux concepteurs d'assistance de spécifier des systèmes d'assistance capables de fournir aux utilisateurs finaux une assistance efficace

pour les aider à réaliser une tâche donnée, en particulier dans le cas de la découverte ou de l'utilisation occasionnelle d'une application-cible.

En exploitant les outils que nous avons développés, nous souhaitons désormais nous intéresser à la définition de systèmes d'assistance adaptés à des applications plus complexes, destinées à un public averti et requérant des connaissances spécifiques.

Par ailleurs, nous souhaitons également faciliter la tâche du concepteur d'assistance. En effet, nous souhaitons aider le concepteur à identifier les besoins d'assistance des utilisateurs finaux de l'application-cible. Pour cela, nous envisageons de proposer au concepteur des indicateurs calculés à partir de traces d'utilisation de l'application-cible. Nous souhaitons également aider le concepteur de l'assistance à améliorer le système d'assistance qu'il a conçu. Pour cela, nous souhaitons notamment permettre aux utilisateurs finaux de l'application-cible d'exprimer leur opinion vis-à-vis de l'assistance qui leur a été proposée. Ces retours d'utilisateurs, accompagnés d'indicateurs calculés à partir de traces d'utilisation de l'application-cible et du système d'assistance, seraient fournis au concepteur de l'assistance afin de lui permettre d'améliorer l'efficacité de son système d'assistance.

Références

DUFRESNE, A. (2001). Conception d'une interface adaptée aux activités de l'éducation à distance - ExploraGraph. In STE, p 301-319.

DUFRESNE, A., BASQUE, J., PAQUETTE, G., LÉONARD, M., LUNDGREN-CAYROL, K. & PROMTEP, S. (2003). Vers un modèle conceptuel générique de système d'assistance pour le téléapprentissage. In Sticef, p 57-88.

CARLIER, F. & RENAULT, V. (2010). Educational webportals augmented by mobile devices with iFrimousse architecture. In International Conference on Advanced Learning Technologies, Tunisia.

GAPENNE, O., LENAY, C. & BOULLIER, D. (2002). Defining categories of the human/technology coupling : theoretical and methodological issues. In ERCIM Workshop on User Interface for All, France, p 197-198.

GINON, B., CHAMPIN, P.-A. & JEAN-DAUBIAS, S. (2013a). Collecting fine-grained use traces in any application without modifying it. In Worshop EXPPORT of ICCBR, New-York, USA.

GINON, B., JEAN-DAUBIAS, S. & CHAMPIN, P.-A. (2013b). Mise en place d'un système d'assistance personnalisée dans une application existante. In IC, Lille, France.

PAQUETTE, G., PACHET, F., GIROUX, S. & GIRARD, J. (1996). EpiTalk, a generic tool for the development of advisor systems. p 349-370.

PAQUETTE, G. (2012). Référencement par compétence, recherche et assistance dans les environnements d'apprentissage et de travail. In TICE, Lyon, France, p 190-199.

RICHARD, B. & TCHOUNIKINE, P. (2004). Une approche centrée modèle pour la construction d'un système conseiller pour un site web. In IC, Lyon, France, p 151-162.

Web social

prototype
organisation
confrontation
forums
évolution
points
communautés
recherche
recommandations
folksonomies
connaissances
santé
nash
scientifique
personnalisées
émotions
construction
information
cible
social
vue
vers
identifier
plus
web
équilibre

Organisation de communautés et Equilibre de Nash

Michel Crampes[1], Michel Plantié[1]

LABORATOIRE LGI2P, Ecole des Mines d'Ales, Site de Nimes
Parc Georges Besse, 30035 Nimes Cedex, France
{michel.crampes,michel.plantie}@mines-ales.fr

Résumé :
La détection de communautés dans les réseaux sociaux est devenue un enjeu majeur pour extraire de la connaissance, observer l'émergence d'opinions ou de savoirs et les réinjecter de manière ciblée sur des groupes sociaux particuliers. Les nombreux algorithmes de détection déjà proposés fournissent des solutions approchées sur laquelle l'analyste ou même les entités regroupées ne peuvent guère intervenir. Dans cet article nous proposons un renversement de perspective. Nous considérons les entités comme des agents qui cherchent à se regrouper et qui peuvent agir sur leur stratégie de regroupement afin d'en tirer le meilleur bénéfice avec différentes règles partagées. Les communautés sont optimisées via la recherche d'un Equilibre de Nash. Cet article explore plus avant ce paradigme en considérant les différentes possibilités offertes par la recherche de l'Equilibre de Nash en matière de réorganisation pilotée par la sémantique, l'intention pragmatique, ou l'évolution spontanée du réseau.

Mots-clés : Equilibre de Nash, Réseaux sociaux, Détection de Communautés.

1 Introduction

La connaissance est indissociable des réseaux sociaux et des communautés. Les opinions, les savoirs et les actions émergent et se structurent au sein de communautés formellement identifiées ou spontanément regroupées. A l'inverse l'observation et l'analyse de regroupements d'entités (classement) font éminemment partie des méthodes de construction de nouveaux savoirs. Les méthodes de classification classiques comme K-Means, l'Analyse en Composantes Principales, la classification hiérarchique, ou la méthode MDS nécessitent de définir à priori le nombre de classes pour certaines, ou un critère d'arrêt portant sur une fonction de coûts. De plus ces méthodes s'appliquent sur des données vectorielles. Les méthodes récentes de détection de communautés ont pour données de départ un graphe uniparti ou un graphe biparti, voire multipartis, en général non orienté. Nous avons montré dans Crampes & Plantié (2013) que la recherche de communautés dans les graphes unipartis, bipartis, orientés ou non, pourait se ramener à la recherche de communautés dans les graphes unipartis.

Les algorithmes de détection de communautés sont pilotés par l'optimisation d'une fonction. Celle qui est la plus considérée est la modularité depuis les travaux de Newman (2006). Elle consiste à considérer un maximum de liens à l'intérieur d'une communauté avec le minimum de liens vers l'extérieur de la communauté. Il est intéressant de noter que certaines méthodes, comme celui de Louvain, considèrent des liens pondérés. A l'inverse des fonctions de clustering ces méthodes ne nécessitent pas de donner en préalable le nombre de communautés ou un seuil sur la fontion. Cependant elles présentent certaines limites qui dans nombre de situations peuvent être gênantes pour une analyse plus sûre et plus détaillée. La première tient au fait que, étant donné le caractère NP-Complet de la recherche de communautés, toutes ces méthodes sont des heuristiques qui fournissent une solution approchée, la modularité maximale n'étant

pas assurée. Dans un article précédent Crampes *et al.* (2013) nous avons montré la possibilité d'améliorer ce résultat en considérant les individus comme des agents engagés dans un jeu non coopératif qui cherchent à optimiser une fonction de gain partagée. Celle-ci, qui est associée à la modularité, nous permet d'introduire une 'fonction potentiel' et de conclure qu'il existe un équilibre de Nash. Autrement dit, partant d'un résultat intermédiaire obtenu par un algorithme de détection de communautés comme par exemple celui de Louvain, il est possible de réaffecter les individus de manière à obtenir un véritable optimum de modularité sans toucher au nombre de communautés déjà trouvées. Ces résultats seront rappelés dans le présent article.

La seconde limite des algorithmes de détection de communautés porte sur le résultat unique du nombre de communautés. Il n'est pas possible de déterminer a priori ce nombre de communautés, ou mieux de réviser le nombre de communautés à partir du résultat final pour obtenir une nouvelle organisation en fonction de nouvelles finalités, ou tout simplement pour chercher à améliorer la modularité. Dans cet article nous montrons comment élargir nos résultats sur l'équilibre de Nash soit pour faire varier le nombre de communautés, soit pour diriger l'organisation des communautés en fonction de la sémantique, celle-ci étant donnée par les proprités des agents dans un graphe biparti.

La section suivante présente un état de l'art de la détection de communautés en mettant en avant les travaux récents qui utilisent l'équilibre de Nash et leur contribution en la matière.

2 État de l'art

La recherche automatique de communautés est devenue un champ de recherche très exploré comme l'atteste l'état de l'art approfondi de Fortunato (2009). La plupart des algorithmes proposent de maximiser une mesure appelée modularité introduite par Newman Newman (2006) et largement appliquée d'abord aux graphes unipartis. Elle a ensuite été étendue aux graphes bipartis dans un premier temps en adaptant sa formulation Murata (2010); Suzuki & Wakita (2009). De nombreuses méthodes de détection de communautés ont été développées à la fois pour les graphes unipartis et bipartis ; le lecteur intéressé trouvera un synthèse dans Crampes & Plantié (2013).

Dans ce même article récent Crampes & Plantié (2013) nous montrons qu'il est possible d'unifier les graphes unipartis, bipartis, et orientés pour produire des communautés à la fois partitionnées et recouvrantes. Nous mettons en œuvre l'algorithme de Louvain Blondel *et al.* (2008) qui fonctionne en agrégeant itérativement les sommets du graphe afin d'augmenter la modularité au maximum. Cependant dans le cas général cet algorithme, comme toute heuristique, ne produit qu'un résultat approché, c'est-à-dire que l'algorithme s'arrête à partir du moment où la modularité ne peut plus augmenter. Or la fonction de recouvrement montre que certains sommets seraient incités à changer de communauté et que ce changement pourrait améliorer ou inversement abaisser le résultat obtenu. Autrement dit le résultat obtenu n'est pas stable. La recherche de communautés stables a fait l'objet de peu de publications, les auteurs se satisfaisant en général des résultats de l'algorithme qu'ils mettent en œuvre et les comparant aux résultats d'autres auteurs. La recherche de stabilité d'un réseau en termes de théorie des jeux a fait l'objet de publications comme par exemple Nisan *et al.* (2007). Selon cette approche la recherche de stabilité de n agents qui choisissent des stratégies à partir de fonctions de satisfaction suppose l'existence d'un Equilibre de Nash. Appliquée à la détection de communautés le problème consiste à trouver les conditions d'existence d'un Equilibre de Nash tel qu'aucun sommet

ne souhaite au final quitter la communauté à laquelle il a été affecté. L'Equilibre de Nash n'a été que peu utilisé pour détecter des communautés. R Narayanam & Y Narahari (2012) l'appliquent sur des graphes unipartis. Ils utilisent la connectivité des sommets pour atteindre un Equilibre de Nash sans mesurer la modularité du résultat. Chen *et al.* (2011) se focalisent également sur les graphes unipartis avec des communautés recouvrantes, et la recherche de l'Equilibre de Nash est l'unique principe directeur. Mais les résultats expérimentaux ne nous semblent pas meilleurs que ceux trouvés par d'autres algorithmes comme celui de Louvain. De plus ces algorithmes ne sont appliqués que sur des graphes unipartis. Dans Crampes *et al.* (2013) nous nous sommes distingués de ces auteurs en recherchant d'abord une solution approchée à l'aide de l'algorithme de Louvain, puis en effectuant des réaffectations pour converger vers un équilibre local qui est prouvé être un Equilibre de Nash. De plus nous avons appliqué cette méthode avec de bons résultats en matière de modularité sur les trois types de graphes. Les deux sections suivantes reprennent synthétiquement ces résultats avant d'introduire des contributions originales.

3 Détection de communautés partitionnées

3.1 Modularité

Le consensus pour la détection de communautés consiste à rechercher une solution approchée qui maximise la modularité selon Newman (2006). Formellement, étant donné un graphe uniparti $G = (N, E)$ représenté par sa matrice d'adjacence A , la modularité Q d'une partition de graphe est définie :

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{1}$$

où A_{ij} représente le poids de la liaison entre i et j, $k_i = \sum_j A_{ij}$ est la somme des poids des arcs attachés au sommet i, c_i est la communauté à laquelle appartient le sommet i, la fonction de Kronecker $\delta(u, v)$ est égale à 1 si $u = v$ et à 0 sinon, enfin $m = 1/2 \sum_{ij} A_{ij}$. Pour l'instant nous ne considérons que des graphes binaires et dans ce cas les poids A_{ij} prennent les valeurs 1 ou 0 selon que la liaison existe ou n'existe pas. L'interprétation de cette formule est la suivante : la modularité est la somme pondérée pour toutes les communautés de la différence entre les liaisons observées à l'intérieur de la communauté (terme A_{ij}) et la probabilité de ces liaisons (terme $\frac{k_i k_j}{2m}$ dont le numérateur est le produit des marges correspondant à la cellule i, j). L'application de cette fonction par de nombreux algorithmes donne de bons résultats. Par exemple elle permet de retrouver des communautés sur des graphes construits ad-hoc. Cependant il a été démontré dans Fortunato (2009) que cette fonction a tendance à fusionner les petites communautés et ainsi à masquer une certaine granularité. Pour les graphes bipartis (et les graphes orientés qui peuvent se ramener à des graphes bipartiis) d'autres formulations ont été proposées. En particulier la formulation dans Barber & Clark (2009) semble faire consensus. Peu différente de celle de Newman, nos travaux récents montrent qu'elle pourrait aussi s'appliquer aux graphes unipartis. Cependant nous utiliserons la formulation de Newman quel que soit le type de graphe pour rester conforme à nos travaux antérieurs publiés afin d'unifier les trois types de graphes. Ayant défini la fonction à optimiser lors de la détection de communautés, nous regardons quel algorithme de détection appliquer.

3.2 Algorithme efficace de détection de communautés partitionnées

Il existe de très nombreuses méthodes de détection de communautés, la plupart optimisant dans la mesure du possible la modularité au sens de Newman pour les graphes unipartis et la modularité au sens de Barber pour les graphes bipartis. Toutes ces méthodes ont en commun d'être des heuristiques puisque la recherche de communautés est par essence un problème NP-complet. Elles fournissent donc des solutions approchées tant en nombre de communautés qu'en matière de répartition des individus entre les communautés. La valeur de la modularité finale n'est pas nécessairement optimale. Dans la section suivante nous présenterons une méthode de réaffectation introduite dans Crampes *et al.* (2013) et basée sur l'Equilibre de Nash pour rechercher à coup sûr un optimum de modularité sans changer le nombre de communautés. Cette méthode, pour l'instant limitée aux liaisons non pondérées, peut s'appliquer au résultat de n'importe quel algorithme de détection de communautés. Nous présenterons ensuite la contribution essentielle de cet article qui portera sur l'application de cette méthode de réaffectation pour introduire des contraintes externes à la création de communauté tant d'un point de vue quantitatif (nombre de communautés, ou nombre d'individus dans les communautés) que qualitatif (répartition des individus, ou répartition des propriétés dans les communautés). Notre approche permet ainsi de manière optimale de passer de la détection de communautés à la construction optimale de communautés sous contraintes, et en particulier de contraintes sémantiques.

Dans les expériences que nous avons menées la méthode que nous utiliserons pour effectuer une première recherche de communauté avant réaffectation est l'algorithme de Louvain Blondel *et al.* (2008) que nous avons présenté dans Crampes & Plantié (2013) et dans Plantié & Crampes (2013). Elle est remarquable par son efficacité et la qualité de ses résultats. Cependant, de nature heuristique elle donne un résultat approché pour lequel la modularité n'est pas optimale. Nous faisons maintenant intervenir une fonction de réaffectation pour rechercher cet optimum.

4 Réaffectation et Equilibre de Nash

L'application d'une méthode de détection comme celle de Louvain nous permet d'obtenir un ensemble C de n communautés. Ces communautés peuvent comporter des individus homogènes (cas d'un graphe uniparti) ou appartenant à deux classes (graphes bipartis). La fonction de réaffectation qui est présentée permet de réaffecter les individus (et les propriétés pour les graphes bipartis) sur les communautés déjà trouvées.

4.1 Fonction de réaffectation RM

Afin de définir cette fonction de réaffectation nous utilisons une variante de l'équation 1 pour les graphes non pondérés. Soit C_i une communauté, $|e_i|$ le nombre d'arêtes dans C_i, d_{C_i} la somme des degrés des noeuds dans $C_i$et m le nombre total de liens dans le graphe. Alors la modularité est :

$$Q = \sum_i \left[\frac{|e_i|}{m} - \frac{(d_{C_i})^2}{(2m)^2} \right] \tag{2}$$

L'interprétation est la suivante : la modularité est la somme pour toutes les communautés du nombre de liens relatifs dans chaque communauté moins la probabilité d'avoir des liens dans cette communauté.

Réaffecter un sommet w de C_1 à C_2 accroit ou décroit la modularité. Nous définissons ce changement comme la mesure de réaffectation de modularité $RM_{w:C_1 \to C_2}$ =$Q_{w \in C_2}$-$Q_{w \in C_1}$ où $Q_{w \in C_k}$ est la valeur de la modularité pour $w \in C_k$ et $C_1 \neq C_2$.

Soit $l_{w|i}$ le nombre d'arêtes entre un sommet w et tous les autres sommets w' tels que $w' \in C_i$, et soit d_w le degré de w. Nous considérons que le sommet w appartenant à C_1 est retiré de cette communauté et ensuite réaffecté à une autre communauté C_2. Nous avons montré dans Crampes *et al.* (2013) que

$$RM_{w:C_1 \to C_2} = \frac{1}{m}(l_{w|2} - l_{w|1}) - \frac{1}{2m^2}[d_w^2 + d_w(d_{C_2} - d_{C_1})] \quad (3)$$

En appliquant l'équation 3, il est possible de vérifier que si nous réaffectons w de C_1 à C_2 puis à nouveau de C_2 à C_1, la modularité ne change pas (le calcul se vérifie de différentes manières) et $RM_{w:C_1 \to C_2 \to C_1} = 0$. En conséquence : $RM_{w:C_1 \to C_2} = -RM_{w:C_2 \to C_1}$.

4.2 Effet de la réaffectation sur les autres sommets

Nous supposons, qu'une première passe de calcul des réaffectation a été effectuée, et à la suite de cela un sommet w du graphe a été déplacé de C_1 à C_2, suite à une mesure de réaffectation positive. Dans notre cas nous choisissons de réaffecter le sommet le plus instable, c'est à dire celui dont la valeur RM est la plus forte. Soit z un autre sommet du graphe. On peut voir la variation de la valeur de réaffectation pour ce sommet z après le déplacement de w.

Le calcul de la différence de la mesure de réaffectation pour le sommet z de l'étape précédente à l'étape actuelle se calcule comme suit :

On cherche $RM^1_{z:C_{from} \to C_{to}} - RM^0_{z:C_{from} \to C_{to}}$ où $RM^0_{z:C_{from} \to C_{to}}$ est la mesure de réaffectation de z avant le déplacement de w et $RM^1_{z:C_{from} \to C_{to}}$ sa mesure après.

Selon les cas de C_{from} et C_{to} on trouve les résultats suivants :

Soit $\triangle R_z = [\{w, z\} - \frac{1}{(2m)} d_z d_w] \frac{1}{m}$, dans lequel $\{w, z\}$ représente le lien entre w et z. S'il n'y a pas de lien cette valeur est nulle.

Dans le tableau ci dessous, nous montrons les différentes valeurs calculées individuellement, de la correction de réaffectation pour z. Dans ce tableau C_3 et C_4 sont des communautés autres que C_1 et C_2.

to\from	C_1	C_2	C_3	C_4
C_1	0	-2$\triangle R_z$	-$\triangle R_z$	-$\triangle R_z$
C_2	2$\triangle R_z$	0	$\triangle R_z$	$\triangle R_z$
C_3	$\triangle R_z$	-$\triangle R_z$	0	0
C_4	$\triangle R_z$	-$\triangle R_z$	0	0

On observe que la matrice est antisymétrique, avec $RM^1_{z:C_{from} \to C_{to}} - RM^0_{z:C_{from} \to C_{to}}$ = -[$RM^1_{z:C_{to} \to C_{from}} - RM^0_{z:C_{to} \to C_{from}}$], ce qui est compatible avec les propriétés de la réaffectation présentées en 4.1.

Ce tableau permet de simplifier les calculs de réaffectation. Il est aussi un moyen intéressant d'étudier l'impact sémantique de la réaffectation. Sans détailler les conclusions dans les limites de cet article, synthétiquement ces propriétés montrent qu'un noeud a tendance à suivre un noeud voisin, ou bien à quitter une communauté dans laquelle arrive un noeud avec lequel il n'a pas de liaison.

4.3 Réaffectation par l'Equilibre de Nash (EN)

Nous montrons dans cette section que l'application de la réaffectation conduit à un Equilibre de Nash, une situation stable où aucun sommet n'a intérêt à quitter la communauté à laquelle il a été finalement affecté. Dans ce but nous ramenons le problème de la réaffectation des sommets à un problème de théorie des jeux non coopératifs. Un lecteur soucieux d'une expression formelle pourra se référer à Crampes *et al.* (2013). En effet les n sommets peuvent être considérés comme n joueurs qui cherchent à optimiser leur gain en jouant des stratégies. Une stratégie est un déplacement accessible à un joueur. Dans notre cas une stratégie est le choix d'une communauté et chaque agent a le choix parmi les m différentes communautés. Un profil de stratégie s à un moment donné est la combinaison des choix stratégiques des agents, c'est-à-dire dans notre cas l'état des affectations à un moment donné.

Une fonction de gain est ce qu'est en droit d'attendre un joueur i en jouant une stratégie compte tenu des stratégies que jouent les autres joueurs. Prenant en compte cette fonction de gain pour chaque joueur, un profil de stratégie s^* est un Equilibre de Nash (EN) si aucun joueur n'a intérêt à changer de stratégie étant donné la stratégie jouée par chacun des autres joueurs. Trois questions se posent : (i) quelles sont les conditions pour qu'il existe au moins un EN, (ii) s'il en existe un comment l'atteindre, et (iii) est-il possible de l'atteindre dans un temps raisonnable polynomial.

Il est possible de donner une réponse à l'existence d'un EN et à la possibilité de converger vers cet équilibre si on peut définir une "fonction potentiel" qui permet d'atteindre un optimum global en recherchant des optima locaux pour les agents. Si de plus la recherche d'un optimum local pour un agent se fait en un temps polynomial, alors le problème entre dans la classe des problèmes PLSComplets (Polynomial Local Search) (E. Tardos & T. Wexler Nisan *et al.* (2007)).

Pour tout jeu fini, une fonction potentiel exacte Φ est une fonction qui fait correspondre à chaque vecteur stratégie s une valeur réelle $\Phi(s)$ avec la condition qu'à une évolution du gain pour un agent correspond la même évolution du gain pour la fonction potentiel. Autrement dit les joueurs partagent un intérêt commun lorsqu'ils choisissent individuellement leur stratégie, même s'ils ne coopèrent pas. L'intérêt de chacun rejoint l'intérêt du groupe. Ainsi armé d'une fonction potentiel il est possible de trouver sûrement un Equilibre de Nash en convergeant à partir de la recherche d'optima locaux pour chaque joueur. Pour trouver de manière convergente un Equilibre de Nash, c'est-à-dire une partition stable des communautés qui satisfasse tout le monde, il nous faut disposer d'une fonction potentiel $\Phi(s)$. La fonction $RM_{w:C_1 \to C_2}$ qui représente le gain f_w attendu par le sommet w quand il est réaffecté est en fait le gain de modularité pour toutes les communautés. Cette fonction $RM_{w:C_1 \to C_2}$ peut en conséquence jouer le rôle de fonction potentiel.

Ce choix de gain local assure donc que l'algorithme local de réaffectation des noeuds présenté dans la section précédente converge bien vers un Equilibre de Nash. Dans la mesure où le calcul de RM est fait à chaque étape de réaffectation uniquement pour les noeuds des deux communautés concernées et que ce calcul est polynomial, l'algorithme de convergence vers l'Equilibre de Nash est PLSComplet (Papadimitriou Nisan *et al.* (2007)).

Il est intéressant de noter que récemment quelques auteurs ont fait appel plus largement à l'Equilibre de Nash pour rechercher les communautés partitionnées R Narayanam & Y Narahari (2012); Chen *et al.* (2011). La méthode que nous avons présentée ici de détection de

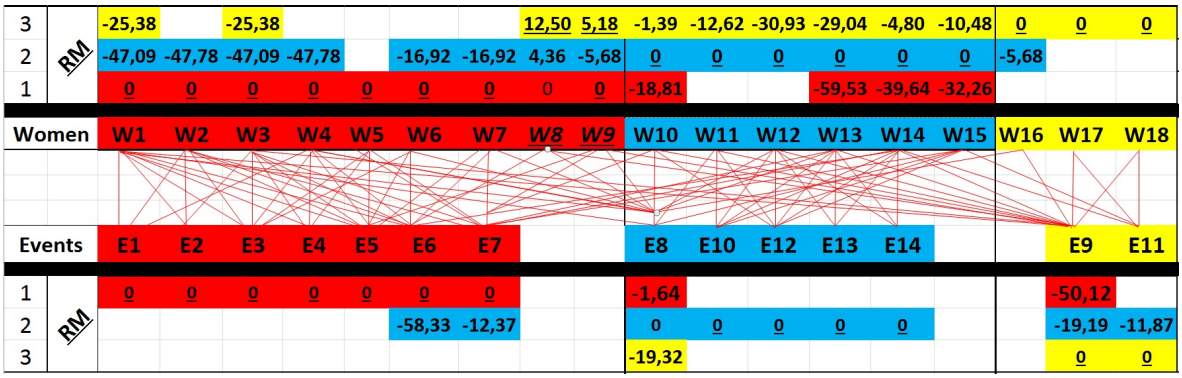


FIGURE 1 – Communautés et mesures de réaffectation pour : Women Events

communautés s'éloigne de ces auteurs parce qu'elle accepte dans un premier temps d'autres heuristiques efficaces bien établies pour rechercher une solution approchée et trouver le nombre de communautés, puis elle finalise la recherche pour atteindre un Equilibre de Nash qui assure la stabilité. A l'inverse les autres auteurs utilisent la recherche de l'Equilibre de Nash pour construire les communautés et réaliser les affectations. De plus leur fonction potentiel diffère de celle que nous avons présentée ici. Nous avons montré dans Crampes *et al.* (2013) que la combinaison Louvain-Réaffectation donne de meilleurs résultats que la recherche de communautés avec uniquement l'Equilibre de Nash. Il est cependant un point essentiel à faire observer. La fonction potentiel que nous utilisons fait que chaque sommet cherche à améliorer la modularité de tout le graphe. Autrement dit nous avons là une fonction socialement très consensuelle dans laquelle l'intérêt de chacun s'identifie à l'intérêt de tous, un peu comme dans un essaim d'abeilles. Cette hypothèse très restrictive sera discutée à la fin de l'article.

4.4 Expérimentations

Des expérimentations de réaffectation ont été réalisées d'abord sur des graphes unipartis, et surtout bipartis de taille petite (de 40 à 1000 noeuds) Crampes *et al.* (2013). L'algorithme de Louvain est appliqué au graphe et quand il s'arrête les RM de tous les noeuds pour toutes les communautés sont calculés. Comme on verra sur l'exemple SW ci-dessous, notre algorithme de réaffectation entre alors en jeu, observe les individus ou les évènements qui ont une propension à changer de communautés (RM positif), réaffecte le plus instable, et recalcule les RM pour les seuls noeuds de la communauté de départ et de la communauté d'arrivée. Ce calcul est réitéré jusqu'à ne plus obtenir de RM positifs. Un équilibre est effectivement atteint assez rapidement dans presque tous les cas expérimentés. Il s'agit d'un Equilibre de Nash où tous les noeuds sont stables (RM négatifs ou nuls). L'algorithme conserve le nombre de communautés trouvées par Louvain et la réaffectation permet d'améliorer la modularité. Sur ce type de graphes les temps de calcul qui dépendent du nombre de communautés sont de l'ordre de quelques secondes. Nous avons réalisé récemment des expérimentations sur des graphes bipartis moyens (environ 10.000 noeuds). Le temps de calcul va jusqu'à 1 minute pour 90 itérations (réaffectations). Pour un graphe de plus grande taille de 120.000 noeuds et 180 communautés trouvées par Louvain qui représente des chercheurs reliés aux articles dont ils sont auteurs le temps d'obtention de l'équilibre est de plus d'une heure dû à un nombre très important de réaffectations, l'heuristique

Individus :

Noeud	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13	I14	I15	I16	I17	I18
1 (7)	0	0	0	0	0	0	0	-10,983	-3,661	-18,811	-41,283	-68,299	-59,525	-39,641	-32,256	-18,432	-29,668	-29,668
2 (6)	-50,625	-50,877	-50,625	-50,877	-41,156	-18,685	-18,685	-7,385	-10,1	0	0	0	0	0	0	-4,797	-27,269	-27,269
3 (2)	-32,445	-46,206	-32,445	-46,206	-32,067	-20,831	-20,831	0	0	-1,389	-12,625	-30,93	-29,037	-4,797	-10,478	0	0	0

Evènements :

Noeud	E1	E2	E3	E4	E5	E6	E7	E8	E10	E12	E13	E14	E9	E11
1 (7)	0	0	0	0	0	0	0	-1,641	-65,964	-79,535	-39,2	-39,2	-84,459	-37,369
2 (5)	-30,678	-30,678	-62,492	-41,156	-73,097	-50,625	-5,555	0	0	0	0	0	-36,359	-10,1
3 (0)	-23,861	-23,861	-48,857	-32,067	-43,681	-32,445	-16,538	-19,316	-44,186	-53,402	-26,133	-26,133	0	0

FIGURE 2 – Equilibre de Nash après réaffectation pour : Women Events

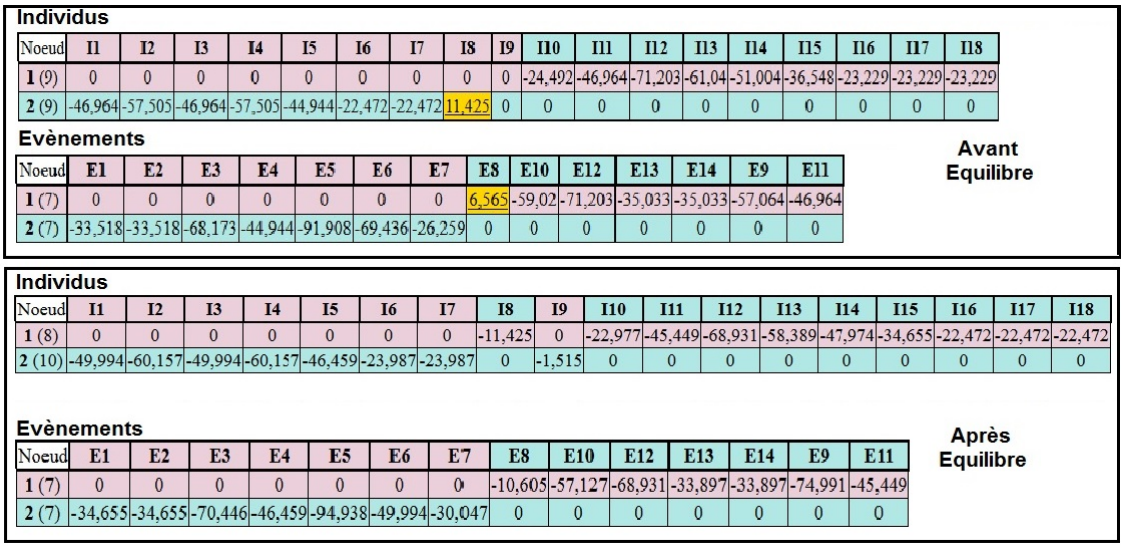
mise en oeuvre par Louvain produisant beaucoup d'instabilités.

Parmi les petits graphes nous détaillons ici celui qui permet de mieux cerner par la suite comment pourront s'appliquer des contraintes. Connu sous le nom Southwern Women (SW) il consiste en un relevé par cinq ethnologues de la partitipation différenciée de 18 dames à 14 évènements sociaux de type "tea party", travail coopératif, etc. à Natchez, Mississipi en 1930. Le but des ethnologues était d'étudier les comportements raciaux au travers des relations individuelles. Formellement il s'agit d'un graphe biparti qui a été intensément étudié et pour lequel beaucoup d'auteurs ont proposé des partitionnements généralement en deux, parfois en trois communautés (voir la meta-analyse de Freeman (2003)).

Nous avons détecté trois communautés regroupant des dames et des événements avec Louvain. Elles sont montrées dans la figure 1 en couleur rouge, bleu et jaune (les individus sont numérotés de I1 à I18, les évènements de E1 à E14). Ce résultat donne une modularité supérieure à tous les autres résultats trouvés dans la littérature et il est donc plus précis, en particulier comparé à ceux présentés dans Freeman (2003). La mesure de RM montre avec accuité la propension à être affecté à une autre communauté. La valeur 0 correspond à la réaffectation dans sa propre communauté. Concrètement deux individus (I8 et I9) sont instables. La figure 2 montre le résultat après réaffectation et obtention de l'Equilibre de Nash. La modularité passe de 0,309 à 0,325. Il est intéressant de noter que les deux individus réaffectés le sont dans la troisième communauté qui se trouve renforcée alors qu'aucun autre auteur n'avait identifié cette communauté. Pourtant c'est cette configuration qui donne la meilleure valeur de modularité.

5 Variation des communautés

Les algorithmes de détection tels que celui de Louvain donnent une solution approchée. De plus le nombre de communautés et l'affectation des noeuds sont des résultats non contrôlés. A l'inverse les méthodes de clustering comme K-means ou l'ACP permettent de controler le nombre de clusters, mais ne permettent pas de faire varier les paramètres autour d'un optimum. Nous montrons ici comment avec la méthode de réaffectation RM-Nash il est possible de faire varier différents paramètres et d'observer les tentatives de retour à l'équilibre. Nous soulignons l'intérêt pratique de ces transformations en particulier au plan sémantique et les champs de recherche qui s'ouvrent.



Individus

Noeud	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13	I14	I15	I16	I17	I18
1 (9)	0	0	0	0	0	0	0	0	0	-24,492	-46,964	-71,203	-61,04	-51,004	-36,548	-23,229	-23,229	-23,229
2 (9)	-46,964	-57,505	-46,964	-57,505	-44,944	-22,472	-22,472	11,425	0	0	0	0	0	0	0	0	0	0

Evènements

Noeud	E1	E2	E3	E4	E5	E6	E7	E8	E10	E12	E13	E14	E9	E11
1 (7)	0	0	0	0	0	0	0	6,565	-59,02	-71,203	-35,033	-35,033	-57,064	-46,964
2 (7)	-33,518	-33,518	-68,173	-44,944	-91,908	-69,436	-26,259	0	0	0	0	0	0	0

Avant Equilibre

Individus

Noeud	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13	I14	I15	I16	I17	I18
1 (8)	0	0	0	0	0	0	0	-11,425	0	-22,977	-45,449	-68,931	-58,389	-47,974	-34,655	-22,472	-22,472	-22,472
2 (10)	-49,994	-60,157	-49,994	-60,157	-46,459	-23,987	-23,987	0	-1,515	0	0	0	0	0	0	0	0	0

Evènements

Noeud	E1	E2	E3	E4	E5	E6	E7	E8	E10	E12	E13	E14	E9	E11
1 (7)	0	0	0	0	0	0	0	-10,605	-57,127	-68,931	-33,897	-33,897	-74,991	-45,449
2 (7)	-34,655	-34,655	-70,446	-46,459	-94,938	-49,994	-30,047	0	0	0	0	0	0	0

Après Equilibre

FIGURE 3 – Réduction de communautés, réaffection et Equilibre de Nash pour : Women Events

5.1 Réduction du nombre de communautés

La première transformation intéressante est de réduire le nombre de communautés à partir d'une structuration intermédiaire, que ce soit le résultat non optimal obtenu par Louvain sur SW ou le résultat optimal obtenu après retraitement par RM-Nash. Ce problème a de nombreuses applications pratiques. C'est en effet celui rencontré par un organe de décision qui souhaite réduire le nombre de sous-structures. Par exemple un laboratoire de recherche comprend trois équipes et on souhaite une meilleure visibilité extérieure en passant à deux équipes plus importantes en taille (le lecteur reconnaîtra un problème familier en général difficile à gérer). L'exemple SW simule ce problème, les chercheurs étant représentés par les dames et les thèmes de recherche, ou les papiers cosignés étant représentés par les évènements.

Pour effectuer une réduction nous opérons de la manière suivante : nous affectons arbitrairement les membres de la communauté qui semble la plus instable (par exemple celle qui présente la plus forte moyenne des RM, ou bien la plus faible numériquement si le but est la visibilité) à l'une des autres communautés. Dans le cas de SW il ne reste plus que deux communautés pour lesquelles tous les RM sont ensuite calculés. Certains noeuds présentent alors des valeurs de RM positives ce qui témoigne de leur instabilité. Nous appliquons ensuite RM-Nash jusqu'à l'équilibre que nous savons exister pour obtenir une affectation sur un nombre de communautés réduit.

Expérimentalement la figure 3 montre le résultat obtenu suite au traitement suivant : on part du résultat non optimal de Louvain qui donne trois communautés, nous fusionnons les communautés 2 et 3 pour ne laisser que deux commautés. Tous les RM sont recalculés sur ces deux communautés. Les noeuds I8 et E8 apparaissent instables avec un RM positif. RM-Nash est ensuite appliqué pour obtenir un équilibre à deux communautés. On peut observer que les deux communautés trouvées sont celles qui sont observées par la majorité des sociologues dans Freeman (2003). Le résultat calculé rejoint donc le consensus. Mais nous pouvons aussi observer que la modularité avec deux communautés est plus faible qu'avec les trois communautés trouvées par Louvain, et a fortiori avec l'optimisation effectuée par RM-Nash comme montré dans

la section précédente. Ainsi nous obtenons bien un équilibre à deux communautés, mais moins bon qu'à trois communautés. En pratique, les agents ainsi réorganisés devraient tisser de nouveaux liens dans leur nouvelle structure pour améliorer l'organisation de l'ensemble. Un point intéressant à noter est que nous obtenons le même résultat expérimental à deux communautés si nous fusionnons au départ les communautés 1 et 2 ou 1 et 3. On peut se poser la question thérorique de savoir si c'est toujours vrai dans le cas général. Nous conjecturons que ce n'est pas le cas car rien n'indique qu'il n'y a qu'un seul équilibre de Nash comme nous allons le voir dans la section suivante.

5.2 Essaimage de communautés et flexibilité structurelle

L'essaimage consiste à rajouter une ou plusieurs nouvelles communautés à celles déjà existantes. Dans une première étape nous partons de n communautés et nous passons à $n+1$. Parmi les applications pratiques on peut citer le désengorgement d'une structure par essaimage, la diversification, etc. Ici aussi les algorithmes de clustering comme K-means nécessitent de recommencer le calcul à partir de zéro sans pouvoir expliciter le phénomène ni justifier un équilibre. Bien évidemment les algorithmes de détection de communautés ne peuvent opérer ce genre de révision puisqu'ils donnent une seule solution. Pour notre part nous partons de n communautés à l'équilibre (cette condition n'est pas cependant nécessaire) et forçons dans une communauté nouvellement créée un sous ensemble d'éléments provenant des autres communautés, En pratique, nous réaffectons l'élément le moins stable même s'il est à l'équilibre (RM négative mais maximale). Cet élément est bloqué dans sa nouvelle communauté et tous les autres RM sont recalculés. L'équilibre est recherché de manière itérative par l'application de notre algorithme RM-Nash. Les résultats expérimentaux sur SW sont montrés dans la figure 4. Nous sommes partis de l'équilibre précédent à deux communautés et avons construit une troisième communauté. L'élément réaffecté en premier est $I9$. Il entraîne avec lui $I7$ et $E7$. L'équilibre est atteint alors avec 3 éléments dans la nouvelle communauté. On se retrouve en final avec une structure à trois communautés très différente de la structure à trois communautés d'où on était parti initialement mais avec une modularité plus faible. Sémantiquement on voit que la nouvelle communauté contient les éléments (les évènements) les plus partagés par les deux communautés initiales. Nous avons réitéré l'expérience en changeant le premier élément forcé et le résultat final varie avec des structures différentes.

Différentes expériences ont aussi été menées pour passer de trois communautés à 4 selon le principe ci-dessus. Plusieurs équilibres ont été trouvés selon le sommet que nous avons privilégié pour ensemencer la communauté rajoutée. Les modularités sont variables dont une encore meilleure que la meilleure à trois communautés. Avec ces expériences il apparait qu'il est possible de faire varier les affectations et le nombre de communautés autour de divers équilibres. Le choix et le nombre des sommets initialement forcés interviennent dans le résultat final et permettent donc de jouer sur la sémantique ou la pragmatique de l'organisation. Notre approche se différencie donc de toutes les autres puisqu'elle permet non seulement de trouver des optima, mais aussi de générer différentes organisations à l'équilibre selon les objectifs de l'expérimentateur.

Nous avons appliqué les mêmes procédures à un autre jeu de données plus conséquent : un graphe biparti de partage de 1000 photos entre environ 300 personnes provenant d'un compte Facebook. Nous avons supprimé une communauté sur la structure trouvée par Louvain. Puis nous avons recherché l'Equilibre de Nash. Enfin comme précédemment nous avons créé une

Individus

Noeud	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13	I14	I15	I16	I17	I18
1 (7)	0	0	0	0	0	0	0	-8,774	-16,16	-30,678	-41,914	-63,628	-63,439	-52,14	-41,472	-20,704	-20,704	-20,704
2 (10)	-57,064	-55,107	-45,828	-55,107	-38,758	-27,522	-16,286	0	-8,964	0	0	0	0	0	0	0	0	0
3 (1)	-41,156	-21,525	-18,685	-21,525	-8,332	-19,568	2,904	-8,648	0	-4,292	-26,764	-40,904	-25,691	-21,841	-11,299	-13,13	-13,13	-13,13

Evènements

Noeud	E1	E2	E3	E4	E5	E6	E7	E8	E10	E12	E13	E14	E9	E11
1 (6)	0	0	0	0	0	0	0,757	-9,469	-52,708	-63,628	-31,246	-31,246	-75,622	-41,914
2 (7)	-37,306	-37,306	-75,748	-49,994	-90,771	-57,064	-20,578	0	0	0	0	0	0	0
3 (1)	-22,914	-22,914	-46,964	-30,804	-41,156	-41,156	0	-12,625	-33,771	-40,904	-19,884	-19,884	-41,409	-26,764

Avant Equilibre

Individus

Noeud	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13	I14	I15	I16	I17	I18
1 (6)	0	0	0	0	0	0	-2,904	-8,774	-14,14	-30,678	-41,914	-63,628	-63,439	-52,14	-41,472	-20,704	-20,704	-20,704
2 (10)	-59,083	-56,874	-47,847	-56,874	-39,768	-28,532	-19,189	0	-7,954	0	0	0	0	0	0	0	0	0
3 (2)	-45,196	-25,06	-22,724	-25,06	-10,352	-21,588	0	-8,648	0	-4,292	-26,764	-40,904	-25,691	-21,841	-11,299	-13,13	-13,13	-13,13

Evènements

Noeud	E1	E2	E3	E4	E5	E6	E7	E8	E10	E12	E13	E14	E9	E11
1 (6)	0	0	0	0	0	0	-16,665	-9,469	-52,708	-63,628	-31,246	-31,246	-75,622	-41,914
2 (7)	-38,063	-38,063	-77,263	-51,004	-81,555	-47,847	-29,289	0	0	0	0	0	0	0
3 (1)	-24,429	-24,429	-49,994	-32,824	-22,724	-22,724	0	-12,625	-33,771	-40,904	-19,884	-19,884	-41,409	-26,764

Après Equilibre

FIGURE 4 – Réduction de communautés, réaffection et Equilibre de Nash puis ajout de communauté et atteinte de l'Equilibre de Nash pour : Women Events

nouvelle communauté à partir de l'élément le plus instable, et nous avons calculé le nouvel Equilibre de Nash. Le premier équilibre est trouvé après 18 itérations. le deuxième équilibre est trouvé après 10 itérations supplémentaires. La modularité sur le deuxième Equilibre de Nash est inférieure à la modularité du premier. Nous retrouvons là les mêmes constatations que ci-dessus.

5.3 Axes de recherche ouverts : sémantique et évolution

Nous décrivons ici quelques pistes exploratoires qu'offre la méthode de variation des équilibres. Nous définissons dans un graphe biparti la sémantique d'une organisation en communautés d'un ensemble d'individus par l'association des éléments de l'autre ensemble. Dans l'exemple SW, les évènements regroupés avec les dames donnent la sémantique des regroupements. Pour un ensemble de chercheurs, les thèmes de recherche donnent la sémantique des équipes auxquelles sont rattachés les chercheurs et les thèmes. La méthode de réduction ou d'extension des communautés présentée ci-dessus permet de rechercher une organisation la plus consensuelle et éventuellement d'effectuer des réorganisations dirigées par la sémantique. En effet en forçant certains éléments sémantiques à être présents dans une même communauté, l'organisation sera sémantiquement forcée. D'autres contraintes peuvent être appliquées comme le nombre minimal ou maximal d'individus dans une communautés. Si ces manipulations conduisent à des équilibres dans le cas de la réduction du nombre de communautés, ce n'est pas toujours le cas pour l'extension pour lequel un équilibre peut ne s'obtenir qu'en vidant une communauté. Les conditions d'obtention d'équilibre sont un axe de recherche prometteur, en particulier en considérant la variation sémantique des communautés : par exemple peut-on rassembler une communauté en équilibre autour d'un sous ensemble de caractérisation des individus ?

Un autre axe de recherche intéressant est celui de l'évolution d'un réseau et de son impact sur l'organisation communautaire. Il fait l'objet de travaux très poussés aujourd'hui comme par exemple dans Roth (2008) parce qu'il permet de mettre en lumière des mouvements sociaux ou

écologiques. La réaffectation peut apporter la possibilité de faire varier certains éléments structurels du réseau (rajout/suppression de sommets, de liaisons) et d'observer l'évolution marginale de l'organisation, voire l'impact de renforcement ou d'affaiblissement de l'organisation.

6 Conclusion

La recherche de communautés dans les réseaux sociaux, et au delà dans les grands graphes a fait l'objet de nombreuses recherches, les plus récentes étant basées sur l'optimisation d'une fonction de modularité représentative d'un découpage 'optimal'. Mais le problème est NP-complet et les algorithmes les plus efficaces sont des heuristiques qui donnent des résultats approchés. En transformant le problème en celui d'un jeu non coopératif, nous avions pu montrer comment améliorer les résultats en recherchant un Equilibre de Nash par réaffectation des sommets. Dans le présent article nous avons poussé plus loin cette logique en donnant un statut d'agents aux sommets et en guidant le découpage dans plusieurs directions (réduction ou augmentation du nombre de communautés, guidage sémantique, évolutivité). Il est ainsi possible de conduire et de moduler l'organisation selon des critères sémantiques ou organisationnels tout en garantissant un optimum. De nombreuses applications sont possibles quand il est question d'analyser des graphes tant dans la sphère sociale que dans des domaines scientifiques (marketing, biologie, médical, etc.) et de faire évoluer des structures.

Références

BARBER M. J. & CLARK J. W. (2009). Detecting network communities by propagating labels under constraints. *Physical Review E - Statistical, Nonlinear and Soft Matter Physics*, **80**(2 Pt 2), 026129.

BLONDEL V. D., GUILLAUME J.-L., LAMBIOTTE R. & LEFEBVRE E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics : Theory and Experiment*, **2008**(10).

CHEN W., LIU Z., SUN X. & WANG Y. (2011). Community detection in social networks through community formation games. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, p. 2576–2581 : AAAI Press.

CRAMPES M. & PLANTIÉ M. (2013). Partition et recouvrement de communautés dans les graphes bipartis, unipartis et orientés. In *IC 2013 Ingénierie des connaissance*.

CRAMPES M., PLANTIÉ M. & LOPEZ M. (2013). Optimisation dans la détection de communautés recouvrantes et équilibre de Nash. In *MARAMI conférence sur les Modèles et l'Analyse des Réseaux : Approches Mathématiques et Informatique*, Saint-Etienne.

FORTUNATO S. (2009). Community detection in graphs. *Physics Reports*, **486**(3-5), 103.

FREEMAN L. C. (2003). Finding social groups : A meta-analysis of the southern women data. In *Dynamic Social Network Modeling and Analysis. The National Academies*, p. 39—-97. Press.

MURATA T. (2010). Detecting communities from tripartite networks. *WWW*, p. 0–1.

NEWMAN M. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E - Statistical, Nonlinear and Soft Matter Physics*, **74**(3 Pt 2), 036104.

NISAN N., ROUGHGARDEN T., TARDOS E. & VAZIRANI V. V. (2007). *[BOOK] Algorithmic game theory*. Cambridde University Press.

PLANTIÉ M. & CRAMPES M. (2013). *Survey on Social Community Detection*. Springer.

R NARAYANAM & Y NARAHARI (2012). A game theory inspired decentralized local information based algorithm for community detection in social graphs . In *ICPR 21st International Conference on Pattern Recognition*, Vienna.

ROTH C. (2008). Coévolution des auteurs et des concepts dans les réseaux épistémiques : le cas de la communauté Â« zebrafish Â». *Revue Française de Sociologie*, **49**(3), 523–553.

SUZUKI K. & WAKITA K. (2009). Extracting Multi-facet Community Structure from Bipartite Networks. *2009 International Conference on Computational Science and Engineering*, **4**, 312–319.

Identifier la cible des émotions dans les forums de santé

Sandra Bringay[1,2], Eric Kergosien[1,3], Pierre Pompidor[1], Pascal Poncelet[1]

[1] LIRMM, UM2-CNRS Montpellier, France
bringay,kergosien, pompidor,poncelet@upmc.fr

[2] AMIS, UM3, Montpellier, France

[3] MAISON DE LA TÉLÉDÉTECTION, IRSTEA, Montpellier, France

Résumé : Les forums de santé en ligne sont des espaces d'échanges où les patients partagent leurs sentiments à propos de leur(s) maladie(s), traitement(s), *etc.* Sous couvert d'anonymat, ils expriment très librement leurs expériences personnelles. Ces forums sont donc une source d'informations très utile pour les professionnels de santé afin de mieux identifier et comprendre les problèmes, les comportements et les sentiments de leurs patients. Dans le cadre du projet Patients' mind, nous nous intéressons à l'analyse semi-automatique de ces forums et plus particulièrement, dans cet article, à la recherche des cibles des émotions exprimées par les patients. Nous proposons une méthode originale pour identifier ces cibles, basée sur la notion de rôles sémantiques et utilisant la ressource FrameNet. Notre méthode a été validée avec succès sur un jeu de données réelles.

Mots-clés : Analyse de sentiments, fouille d'opinion, cible des émotions, forums de santé.

1 Introduction

Dans le cadre du projet Patients' mind [1], nous nous intéressons à l'analyse semi-automatique des forums de santé en ligne. Ces forums sont des espaces d'échanges où les patients, sous couvert d'anonymat, relatent très librement leurs expériences personnelles. Citons, par exemple, les forums très actifs tels que *Doctissimo, Same-story* ou *Allo Docteurs* [2], qui permettent à des internautes (souvent non professionnels de santé) d'échanger à propos de leur santé. Hancock *et al.* (2007) ont montré que la communication et l'anonymat derrière un ordinateur facilitent l'expression d'états affectifs comme les émotions, les opinions, les doutes, ou les évocations de risques, qui sont généralement réprimés lors de communications plus traditionnelles comme des interviews en face à face ou des réponses à des enquêtes. Ces ressources s'avèrent donc très riches pour les professionnels de santé qui ont accès à des échanges entre patients, entre patients et professionnels et même entre professionnels.

Dans la littérature, de nombreux travaux traitent de l'analyse des sentiments. Beaucoup portent sur des tâches de classification selon la polarité (*positif, négatif* ou *neutre*) (Pang *et al.*, 2002), l'émotion (*joie, colère, etc.*) (Strapparava & Mihalcea, 2008a) ou encore l'intensité de ces deux états affectifs (Wiebe, 2000). Ces approches ont été appliquées dans des domaines variés et permettent de classer efficacement un texte. Cependant, elles s'intéressent rarement à la détection des cibles ou des sources alors que ces dernières sont porteuses de nombreuses informations. Considérons par exemple les deux phrases suivantes : « *J'ai peur de la réaction de mon médecin* » et « *Mon médecin a peur de ma réaction au médicament* ». Les méthodes

1. Financé par la MSH-M (Maison des Sciences et de l'Homme de Montpellier) et le réseau inter-MSH https ://www.lirmm.fr/patient-mind/

2. http ://www.doctissimo.fr/, http ://www.same-story.com/, http ://www.allodocteurs.fr

classiques vont détecter que ces deux phrases sont *négatives* et qu'elles contiennent l'émotion *peur*. Dans le premier cas, elles ne détecteront pas que l'émotion porte sur la réaction du médecin et qu'elle est ressentie par le locuteur. Dans le deuxième cas, elles ne détecteront pas que l'émotion porte sur la réaction du patient au médicament et qu'elle est ressentie par le médecin. C'est ce niveau de précision qui nous intéresse dans ces travaux pour pouvoir ensuite agréger ces informations pour plusieurs messages (*e.g.,* trouver le nombre de messages dans lesquels des patients expriment leur peur à propos d'un médicament). Dans la littérature, on distingue des méthodes qui analysent la source (*e.g.,* qui ressent l'état affectif ?) (Bethard *et al.*, 2004; Choi *et al.*, 2005) ou la cible de l'état affectif (*e.g.,* sur quoi porte l'état affectif ?) (Wu *et al.*, 2009). Dans cet article, nous nous focalisons sur la cible, dans le contexte spécifique des émotions exprimées dans les forums de santé.

Notre méthode vise à expliciter des traces d'émotions en identifiant dans les textes une cible ou un contexte, si possible médical, facilitant l'interprétation. Cette méthode pourra être généralisée à d'autres états affectifs et à d'autres domaines d'application. Pour cela, nous proposons d'intégrer l'analyse sémantique de surface (*Shallow semantic parsing*) et utilisons la ressource lexicale FrameNet [3]. Basée sur la notion de rôle sémantique définie par (Baker *et al.*, 1998), elle permet de décrire schématiquement des situations grâce à un système relationnel de concepts (quels éléments dans la phrase participent, subissent, causent, etc. une situation ?). Une annotation basée sur cette ressource nous permet de repérer dans les phrases des expressions d'émotions et d'en expliciter les constituants. Nous proposons une typologie des annotations dédiées à notre contexte d'étude spécifique. À notre connaissance, il n'existe pas de méthode basée sur ce type d'annotations et personnalisée pour les forums de santé. Nous comparons cette approche à celle plus classique qui se base sur un calcul de distance dans l'arbre syntaxique pour identifier des associations entre émotions et des cibles prédéfinies. Notre méthode a été expérimentée avec succès sur un jeu de données réelles et validée par 10 annotateurs humains, tous chercheurs en informatique.

Le reste de ce document est organisé comme suit : dans la section 2, nous motivons nos travaux dans le cadre de l'analyse semi-automatique des forums de santé et nous définissons la tâche d'analyse de sentiments visée. Dans la section 3, nous présentons les méthodes récentes dédiées à cette tâche. Dans la section 4, nous décrivons les expérimentations menées ainsi que les principaux résultats obtenus dans la section 5. Enfin, dans la section 6, nous concluons et donnons les principales perspectives associées à ces travaux.

2 Motivations et définition de la tâche visée

Comme souligné par (Siegrist, 1994), l'un des grands challenges de la santé de demain est de capter la satisfaction des patients, devenus des clients, pour répondre à la question : *Comment s'améliorer ?* Avec cet objectif, il a étudié les commentaires des patients à l'issu de séjours hospitaliers pour les transformer en données manipulables par les autorités médicales pour la prise de décisions. En utilisant les forums comme objet d'étude, nous franchissons une étape supplémentaire, en approchant la *sphère privée* du patient. En effet, ce dernier s'exprime dans les messages tel qu'il ne le ferait pas via des questionnaires, même anonymes. Toutefois, identifier précisément l'état affectif des patients via ces messages reste difficilement analysable objecti-

3. https ://FrameNet.icsi.berkeley.edu/fndrupal/home

vement et surtout vérifiable (Quirk, 1985). On peut néanmoins envisager d'utiliser ces grandes quantités de textes pour construire des indicateurs qui soient pertinents pour les professionnels de santé. Un exemple d'une telle application est *We feel fine* [4] qui parcourt le web pour prendre la température des humeurs des internautes. Toutes les 10 minutes, elle récolte des phrases contenant des mots d'émotions puis réalise des calculs statistiques par type de sentiments, âge, *etc.* Dans le contexte éminemment subjectif des forums de santé, la caractérisation et la compréhension de ces états affectifs est difficile mais néanmoins particulièrement intéressante dans la perspective de compléter et d'améliorer les programmes de santé publique, notamment quand on peut associer ces états affectifs à des cibles médicales précises. Un exemple d'application s'adresse aux laboratoires pharmaceutiques qui dans une perspective de veille informationnelle, cherchent à identifier les forums dans lesquels les patients s'expriment à propos de leurs médicaments et les états affectifs associés. Ce feedback peut les aider à améliorer leurs produits ou leur communication à propos de ces produits. Un autre exemple est celui des médecins qui veulent connaître les craintes des patients vis à vis des traitements prescrits. Ce feedback peut les aider à améliorer les communications faites aux patients.

Depuis le début des années 2000, l'analyse de sentiments, également appelée fouille d'opinions (*opinion mining*), a connu un intérêt croissant. Beaucoup de communautés se sont intéressées à ce domaine et ont donné des définitions et interprétations variées (*e.g.,* psychologie, sciences sociales, linguistique computationnelle, traitement automatique du langage, fouille de données, *etc.*). En fait, l'analyse de sentiments vise l'extraction des états affectifs exprimés explicitement ou implicitement dans des textes. On trouve plusieurs modélisations de l'opinion (Kim & Hovy, 2004; Kobayashi *et al.*, 2007), qui diffèrent selon l'objet de l'étude et les tâches réalisées à partir de ces modèles. Afin de généraliser nos travaux à n'importe quel état affectif, nous considérerons la définition suivante. Un *état affectif* est ressenti par un *émetteur* (ou *source*). Il fait référence à une *polarité*, c'est-à-dire un jugement pouvant être *positif* s'il est lié à un effet bénéfique sur l'émetteur, *négatif* dans le cas contraire ou *neutre*. L'état affectif peut également faire référence à une *émotion* comme la *colère*, la *joie*, la *tristesse*, *etc.* Généralement, les émotions sont associées à une polarité. La *joie* est considérée par exemple comme positive et la *colère* comme négative. On peut associer différents niveaux d'*intensité* à l'état affectif (*e.g., très positif, un peu triste, etc.*). Pour finir, l'état affectif porte sur une *cible* qui est le réceptacle de la polarité ou de l'émotion. Par exemple, dans la phrase « *Je suis très contente de la chirurgie* », l'émetteur est l'énociateur, la polarité est *positive*, l'émotion est la *joie*, l'intensité est *élevée* (présence du mot modifieur « *très* ») et la cible est la « *chirurgie* ».

Dans ce travail, nous nous focalisons sur l'identification des cibles. Elles sont généralement présentes dans les textes sous la forme d'Entités Nommées (EN), d'événements, de concepts abstraits, de caractéristiques associées à ces concepts abstraits ou de contextes généraux traités par l'état affectif (Popescu & Etzioni, 2005; Wilson *et al.*, 2005). Considérons les exemples décrits dans la Table 1. Dans les phrases P1, P2 et P3, l'émotion *peur* porte sur une entité représentée respectivement par le concept général « *médicament* », l'événement « *début de la chimiothérapie* » et l'EN « *IVEMEND* » (qui est un nom de médicament). Dans la phrase P4, la cible de la polarité est plus complexe et porte sur un *aspect*, une caractéristique : « *le taux de tolérance* » de l'entité « *médicament* ». Dans la phrase P5, seul l'aspect est présent. Dans la phrase P6, il n'y a pas de cible explicite. L'émotion fait référence au contexte général dans

4. http ://www.wefeelfine.org/

lequel la phrase est énoncée. Dans la phrase P7, la cible est détaillée dans le reste de la phrase et ne se limite pas à l'entité médicale « *douleur* ». Il est parfois difficile de distinguer la cible des circonstances ou causes ayant suscité l'état affectif. C'est le cas par exemple dans la phrase P8. Ces exemples illustrent la complexité de la tâche visée, consistant à identifier des cibles pouvant s'exprimer de manières très variées. Contrairement à la plupart des approches de la littérature, nous prenons le parti pris dans cette proposition de ne pas limiter la cible à quelques mots mais au contraire de proposer le plus d'informations possibles comme dans les phrases P7 et P8.

P1 : J'ai **peur** de ce médicament.
P2 : J'ai **peur** de commencer la chimiothérapie.
P3 : J'ai **peur** de prendre de l'IVEMEND.
P4 : Le taux de tolérance pour la moyenne des patients pour ce médicament est **excellent** !
P5 : Son taux de tolérance est **excellent**.
P6 : J'ai **peur**.
P7 : J'ai **peur** de vivre dans la douleur encore une dizaine d'années.
P8 : On m'a donné une chance sur 10 de vivre pendant 10 ans, annoncé de réelles perspectives de récurrence et la **peur** est restée avec moi pendant tout ce temps.

TABLE 1 – Exemple d'expressions de l'émotion.

D'un point de vue technologique, l'analyse (semi-)automatique des forums est difficile et cela est également vrai pour la tâche visée de recherche des cibles des émotions. La plupart des méthodes (semi-automatiques) utilisées pour le domaine de la santé ont été appliquées sur des publications, des comptes rendus d'hospitalisation, *etc.* L'adaptation de ces méthodes aux textes issus des médias sociaux comme les forums est loin d'être triviale. En effet, les messages sont écrits par les patients, de manière peu rigoureuse. Ils sont de taille variable (entre une centaine de caractères et un millier). Ils contiennent des structures grammaticales non conformes, de nombreuses fautes d'orthographe, des abréviations, des expressions porteuses de sentiments comme des mots d'émotions (« *j'aime* » ou « *je déteste* »), des mises en forme particulières (lettres capitales « *PLUS* », répétées « *ASSSSEEEZ* », suite de ponctuation répétées « *! ! !* »), des mots d'argot spécifiques ou non aux thèmes des forums (« *LOL* », « *FIV* ») et des émoticons (« *:-)* »). Le volume des messages est généralement très important (dans le forum réservé au cancer du sein du site Doctissimo [5], on trouve plus de 3300 discussions dont certaines contiennent plus de 2000 réponses). Finalement, traiter les forums de santé avec des méthodes semi-automatiques pour en extraire de l'information reste un véritable challenge. La méthode proposée dans ce papier est efficace sans prétraitement spécifique comme un correcteur orthographique ou grammatical qui est difficile à implémenter de manière générique pour les forums de santé à cause du vocabulaire non standardisé qui varie beaucoup d'un forum à l'autre.

3 Etat de l'art

Lorsque l'on cherche à associer des états affectifs à des cibles, avant d'identifier les liens, une première étape consiste à repérer dans les textes des candidats pour ces deux catégories.

5. http ://forum.doctissimo.fr/sante/cancer-localisation/Sein/liste_sujet-1.htm

Pour identifier les marques d'émotions, il existe de très nombreuses ressources (liste de mots, phrases, idiomes). La plupart ont été construites pour l'anglais et l'analyse de la polarité : General Inquirer (Stone & Hunt, 1963), Linguistic Inquiry and Word Count (Tausczik & Pennebaker, 2010), MicroWNOp (Cerini *et al.*, 2007). Toutefois, des ressources plus spécifiques, comme le dictionnaire DAL sentiment dictionary (Whissell, 1989) ou le lexique de (Mohammad & Turney, 2010) ont été créées pour les mots chargés d'émotions. On trouve également des approches qui cherchent à étendre ces vocabulaires pour des domaines spécifiques en construisant des règles manuelles (Neviarouskaya *et al.*, 2011), en trouvant des mots co-occurrants à partir de mots déjà identifiés comme dénotant des états affectifs, via des corpus volumineux ou le web (Harb *et al.*, 2008; Kozareva *et al.*, 2007). On trouve également des méthodes qui ne se limitent pas à l'utilisation de lexiques comme (Strapparava & Mihalcea, 2008b) qui implémentent des approches de machine learning. **Pour identifier les cibles potentielles des états affectifs,** les approches sont en général spécifiques au domaine d'application. (Hu & Liu, 2004) ont utilisé un algorithme de règles d'association pour identifier les caractéristiques fréquentes dans les avis sur les produits. (Zhuang *et al.*, 2006) ont utilisé pour les critiques de films, des données étiquetées et des patrons grammaticaux comme modèle. Dans le cas des forums de santé, l'identification des cibles est plus difficile car les auteurs évoquent de nombreuses entités, difficiles à comparer et lister *a priori* comme nous l'avons montré dans les exemples de la Table 1.

Une fois les candidats représentant les états affectifs et les cibles générés, trois grandes familles d'approches permettent de les relier. **La première famille de méthodes reliant états affectifs et cibles** prend en considération des aspects essentiellement linguistiques, représentés sous forme de règles (Hu & Liu, 2004; Mudinas *et al.*, 2012) comme les inverseurs de polarité (*sentiment ou valence shifters, e.g., ne pas, à peine, très, etc.*) ou les conjonctions (*e.g.,* le médicament x est bien *mais* Y est le meilleur). Ces règles sont très complexes et certaines ont été théorisées dans le cadre de l'étude de la *Sémantique compositionnelle* (Dowty *et al.*, 1989) qui considère que la signification d'une expression est fonction de la signification de ses constituants et de règles de composition. Par exemple, « *ma douleur a été réduite significativement* » est une expression positive, composée d'un élément négatif (« *douleur* ») et d'une relation (« *réduite significativement* »). L'efficacité de cette première famille de méthodes est fortement liée aux style de langue qui impacte sur les règles linguistiques à prendre en compte. Dans le contexte des forums de santé, elles sont difficiles à mettre en œuvre car ces styles varient d'un forum à l'autre. Bien qu'il existe des exceptions, dans les forums sur la maternité, s'adressant à des jeunes femmes, le registre de langue est souvent très familier, proche de l'écriture SMS, alors que le registre dans les forums traitant de la douleur du dos et s'adressant essentiellement à des personnes âgées est beaucoup plus soutenu. Pour notre domaine d'application, il est donc assez difficile de mettre en place une méthode générique à tout type de forums, en s'appuyant uniquement sur des approches basées sur les lexiques et des règles. **La seconde famille de méthodes** se base sur différents calculs de distance entre les mots dénotant les états affectifs et les cibles potentielles. La plus couramment utilisée est la proximité : l'expression de l'état affectif retenue est la plus proche de la cible en nombre de mots (Hu & Liu, 2004). Il est également possible d'utiliser l'arbre syntaxique d'une phrase déterminant les dépendances. Par exemple, Zhuang *et al.* (2006) considèrent cet arbre comme un graphe et calcule la distance entre une cible candidate et un état affectif candidat par un parcours en largeur où le plus court chemin est calculé en nombre d'arcs. Wu *et al.* (2009) considèrent plutôt la distance en profondeur entre la cible candidate et l'état affectif candidat en recherchant le plus petit parent commun

dans l'arbre syntaxique. Pour améliorer les performances de ces méthodes, **une troisème famille de méthodes** propose de combiner une approche linguistique à une approche prenant en compte la distance entre les mots. Ces méthodes, dites hybrides, introduisent des connaissances linguistiques pour pondérer les arcs entre les nœuds de l'arbre syntaxique (Ding & Liu, 2007). Dans notre contexte, à cause des particularités du langage, nous montrerons l'importance de la robustesse du parseur utilisé.

Dans cet article, nous proposons d'intégrer l'analyse sémantique de surface (*Shallow semantic parsing*), s'apparentant à la première famille de méthodes, pour non seulement identifier des traces d'émotions dans le forums de santé mais également améliorer la recherche des cibles associées à ces émotions. Il s'agit d'associer des rôles sémantiques aux différents constituants d'un énoncé, c'est-à-dire une fonction par rapport à une situation schématique de type *Agent, Patient, Sujet*, *etc.* Nous nous sommes intéressés tout particulièrement à la ressource développée dans le cadre du projet FrameNet (Baker *et al.*, 1998). Ces auteurs définissent des *Frames*, correspondant à des représentations schématiques de situations. Les rôles sémantiques, appelés *Frame Eléments* (FE), sont associés de manière unique à ces frames. Ils sont exprimés par des *Unités Lexicales* (UL). Le tableau 2 décrit un exemple simplifié des informations contenues dans la base FrameNet pour la frame EXPERIENCER FOCUS. Cette frame possède deux FEs principaux : l'*expérimentateur* et la *cible*. D'autres FEs peuvent lui être rattachés, comme les *circonstances*. Comme le montrent les exemples d'annotations, les mêmes FEs peuvent être évoqués par des constituants de natures syntaxiques et grammaticales différentes.

Définition des FEs : Cette frame décrit les EMOTIONS d'un **expérimentateur** à l'égard d'une cible. Des *circonstances* liées à l'émotion peuvent également être exprimées.
Exemples d'annotations : **My** ENJOYMENT of the movie was considerably impaired by the seven-doot guy sitting in front of me. **Smith** takes great PLEASURE in collecting matchboxes. **I** HATE you *when you do that*.
Exemples d'ULs : abhor.v, abhorrence.n, abominate.v, adoration.n, adore.v, afraid.a, agape.a, antipathy.n, apprehensive.a, calm.a, comfort.n, compassion.n, cool.a, delight.v

TABLE 2 – Exemple simplifié extrait de la frame Experiencer Focus.

Cette théorie a déjà été appliquée avec succès à la traduction automatique (Boas, 2002) et aux systèmes de question/réponse (Narayanan & Harabagiu, 2004). À notre connaissance, seul (Kim & Hovy, 2006) ont utilisé la ressource FrameNet pour l'identification des états affectifs. Dans cet article, nous nous proposons de traiter spécifiquement le cas des émotions et nous affinons la notion de cible utilisée par ces auteurs en distinguant différents types de frames pour expliciter au mieux les états affectifs exprimés en relation avec le domaine de la santé.

4 Premières évaluations

La figure 1 décrit la démarche globale visant à identifier des états affectifs et des cibles. Elle se décompose en trois étapes :

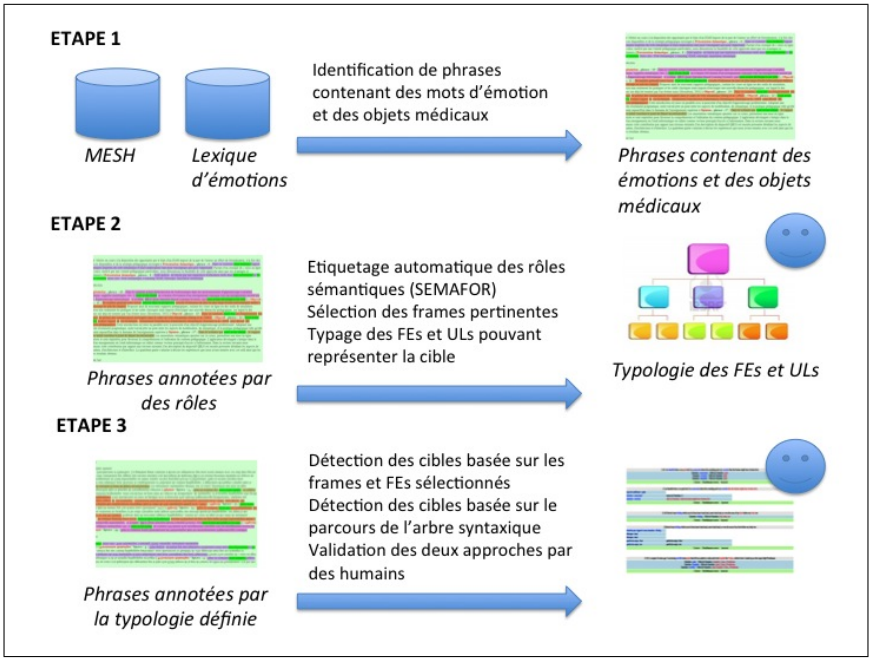


FIGURE 1 – Démarche globale de notre approche.

Etape 1 : Constitution du corpus. Nous avons construit un corpus à partir de 17000 messages issus du forum de santé anglais *Spine-health* [6]. Nous avons annoté automatiquement ce corpus avec le lexique d'émotions de (Mohammad et Turney, 2010). Ce lexique est composé de plus de 14000 entrées caractérisées par leur polarité et associées à 8 émotions. Dans ce travail, nous nous intéressons uniquement aux six émotions de (Ekman, 1992) : *colère, dégoût, peur, joie, tristesse, surprise*. Cette annotation automatique permet de filtrer les messages objectifs (sans état affectif), soit 22% des messages. Afin de ne travailler que sur des émotions portant sur des objets médicaux, nous avons utilisé le MeSH [7] pour repérer des entités médicales, ce qui nous a permis de filtrer les messages n'en contenant pas, soit 6% des messages. Dans un message, plusieurs émotions sont généralement exprimées du fait de leur longueur relative. Nous avons donc choisi de segmenter les messages en phrases. Nous avons finalement gardé 1000 phrases pour constituer notre corpus d'étude.

Etape 2 : Identification des frames pertinentes pour la détection de cibles d'émotions. Nous avons utilisé l'outil SEMAFOR (Das *et al.*, 2010) qui annote des textes selon les éléments des frames de la ressource FrameNet. Un exemple de phrase annotée est « I[expérimentateur]'m afraid[EMOTION] of getting an addiction[Cible] when I start X[CIRCONSTANCE] ». Nous avons étudié manuellement les frames repérées par cet outil et identifié celles qui sont pertinentes pour la recherche des cibles des émotions dans le contexte médical. Parmi les 1164 frames existantes, nous en avons sélectionné 16 comme relatives à l'expression des états affectifs, 5 explicitant de manière générale l'expression d'une émotion et 7 spécifiques aux objets

6. http ://www.spine-health.com/forum
7. http ://www.ncbi.nlm.nih.gov/

médicaux. Pour chaque frame, nous avons choisi les FEs pouvant jouer le rôle d'émotion ou de cible. Par exemple, pour la frame FEAR, nous avons utilisé le FE *expressor* comme émotion et les FEs *topic* et *stimulus* comme cible. Le tableau 3 liste ces frames. Si on reprend l'exemple précédent, la cible sera étiquetée par « Experience bodily harm ».

Traces d'émotions	**Cibles explicitant la trace de l'émotion**
Experiencer focus	**Contextes Généraux**
Experiencer obj	Mental stimulus exp focus
Emotions	Partiality
Emotion active	Activity start
Emotion directed	Causation
Emotion heat	Awareness
Emotions by stimulus	
Emotions success or failure	**Contexte médical**
Complaining	Medical conditions
Contrition	Observable body parts
Desirability	Perception body
Desiring : Event	Cause harm
Fear	Intoxicants
Feeling	Cure
Sensation	Experience bodily harm
Tolerating	

TABLE 3 – Typologie des frames d'intérêt pour la recherche des cibles médicales des émotions identifiées manuellement.

Etape 3 : Evaluation des cibles identifiées. Dans le corpus initial, nous n'avons retenu que les phrases correspondant aux frames sélectionnées à la deuxième étape (soit 345 phrases). Nous avons mis en forme la sortie de l'outil SEMAFOR selon la typologie décrite dans le tableau 3. Nous utilisons une couleur de caractères pour les informations relatives aux émotions et une autre pour celles liées au contexte et pouvant aider à l'interprétation des cibles associées à ces émotions. La sortie présentée à l'utilisateur correspond donc à une liste d'informations typées. Nous notons cette approche l'*Approche Rôles*. Nous avons également proposé pour chaque phrase des liens entre émotions et objets médicaux basés sur le calcul de distance dans l'arbre syntaxique. Pour cela, nous avons utilisé l'étiqueteur morpho-syntaxique développé à Stanford [8]. Pour déterminer la distance entre émotions candidates et cibles candidates, nous avons choisi le calcul basé sur le plus court chemin en nombre d'arcs qui s'était avéré être le plus efficace dans la littérature (voir Section 3). La sortie présentée à l'utilisateur correspond à des couples *Emotion - Entité médicale*. Nous notons cette approche l'*Approche Chemin*. Les résultats de ces deux approches ont été validés via une interface web.

Nous avons déjà montré dans un travail précédent (Melzi *et al.*, 2014), la difficulté pour un humain d'identifier une émotion. Cela est d'autant plus vrai lorsque l'on essaie d'identifier la cible associée à une émotion. Dans le but de valider les informations obtenues avec les deux

8. http ://nlp.stanford.edu/software/stanford-dependencies.shtml

approches, nous avons demandé à 10 chercheurs en informatique si les informations obtenues étaient : *Correctes*, *Partiellement correctes* (si une partie de la cible avait été oubliée ou ajoutée) ou *Incorrectes*. Au moins trois experts ont répondu à cette question pour chaque phrase.

5 Résultats des expérimentations et discussions

Le tableau 4 présente les résultats obtenus à partir de la méthodologie décrite dans la Section 4. Nous avons étudié l'accord entre annotateurs en utilisant la mesure $Kappa\ Fleiss$ et une mesure basée sur les classes d'équivalence régulières (CER) qui prend en compte la gradation dans les notations (c'est-à-dire le fait que la note *correcte* est plus proche de la note *partiellement correcte* que *incorrecte*). Dans les deux cas, nous avons obtenu un *accord modéré*. Il est important de noter que, pour 87% des phrases, l'information apportée par l'*Approche Rôle* a été jugée *correcte* ou *partiellement correcte*.

	Correct	Partiellement correct	Incorrect	Kappa Fleiss	CER
Approche Chemin	8%	43%	49%	0.48 (Moderate)	0.52
Approche Rôles	53%	34%	13%	0.58 (Moderate)	0.63

TABLE 4 – Résultats

Comme l'on pouvait s'y attendre, l'*Approche Chemin* est la moins efficace. La raison principale est liée au fait que dans notre cas d'étude et contrairement à d'autres domaines d'application, il n'est pas possible ici de définir à l'avance une liste de cibles exhaustives. Par exemple, pour l'étude des avis sur les produits, les cibles sont liées aux produits eux mêmes ou à leurs caractéristiques et il est possible de les identifier facilement car elles apparaissent fréquemment dans les commentaires. Dans les forums, les cibles sont très diversifiées et elles ne sont pas toujours limitées aux objets médicaux que l'on repère en utilisant le MeSH. Par exemple, dans la phrase « I fear the long term tendency », l'*Approche Chemin* va identifier « fear » comme étant l'émotion mais ne sera pas capable de l'associer à « tendency » qui n'est pas repérée comme entité médicale. Une validation de l'étape d'annotation est donc également nécessaire. Par ailleurs, les émotions sont dépendantes de la sensibilité du patient et il est très subjectif de distinguer dans la phrase précédente la peur d'une manisfestation de l'incertitude sur les effets d'un produit ou d'une imprécision sur le diagnostique. Une deuxième limitation est liée au peu de performance du parseur utilisé pour extraire l'arbre syntaxique sur les phrases des forums qui sont très souvent mal construites (problème de ponctuation, d'orthographe, *etc.*). Le parseur SEMAFOR utilisé dans l'*Approche Rôles* semble plus robuste aux spécifités du langage. Pour finir, l'étude des phrases exclues à l'étape 2 (et qui n'auraient pas été exclues si l'on avait utilisé uniquement l'*Approche Chemin*), nous a montré que la plupart des phrases avaient été sélectionnées dans le corpus initial du fait de la présence d'un mot d'émotion, qui n'est pourtant pas toujours représentatif de l'expression d'un sentiment dans le contexte des forums de santé. Par exemple, la phrase « *If you want your curve progression halted, it can only be done by surgery* » a été retenue à cause de la présence du mot « *progression* » se trouvant dans la ressource généraliste des émotions et qui, dans notre contexte, ne représente pas une réèlle émotion.

L*'Approche Rôles* est la plus efficace notamment quand l'émotion est portée par un verbe (e.g., « *I fear the surgeon will be reluctant to continue helping control my pain* »). Elle montre

toutefois certaines limites. L'étude des phrases ayant sucité des écarts d'évaluation nous a permis de montrer que les cibles identifiées comme étant relatives au domaine de la santé sont plus faciles à interpréter que les cibles générales. Par exemple, « *I hope that your injection starts to bring you relief soon* » est simple à analyser alors que la phrase « *I call the hospital all the time and they say not to worry* » a généré des écarts d'annotations. Par ailleurs, la ressource FrameNet n'intègre pas aussi exhaustivement tous les états affectifs que l'on voudrait repérer dans les textes. Par exemple, seules deux frames permettent d'identifier l'incertitude (*Frame Certainty, Degree*) qui est particulièrement intéressante à prendre en compte dans les forums de santé. Pour finir, la généralisation de cette méthode à d'autres langues est difficile. Il n'existe par exemple pas de ressource aussi complète que FrameNet pour le Français. Une perspective évidente consiste à reproduire cette étude sur d'autres forums.

6 Conclusions et perspectives

Les forums représentent une base volumineuse et variée des connaissances et des perceptions qu'ont les patients de leur maladie et des soins qui leur sont éventuellement prodigués. Dans cet article, nous avons décrit une approche permettant d'aider un lecteur à identifier des traces d'émotions dans les messages des forums de santé et d'en expliciter les constituants. Nous avons montré que l'utilisation d'un étiqueteur de rôles sémantiques s'avérait tout à fait efficace pour interpréter les cibles de ces émotions sans aucun prétraitement des messages.

Les perspectives associées à ce travail sont nombreuses. Tout d'abord, nous nous sommes limité aux cibles présentes dans les phrases mais nous n'avons pas travaillé sur les relations inter-phrastiques au niveau du paragraphe ou du message. Par ailleurs, identifier le porteur de l'émotion pourrait représenter une information supplémentaire tout aussi pertinente pour l'analyse des forums et il s'agirait d'une extension très simple à réaliser car cette information est déjà présente dans les éléments retournés par l'étiqueteur de rôles sémantiques. En effet, contrairement à l'analyse des avis de produits qui contiennent généralement uniquement les sentiments des auteurs des commentaires, les patients dans les forums relatent des émotions qui ne leur sont pas propre (*e.g.,* « *mon médecin s'inquiète de voir mon taux de glycérol augmenter* »). De plus, une fois ces relations entre cible et émotion identifiées, elles peuvent être généralisées pour un ensemble de messages afin de résumer les états affectifs de différents émetteurs à propos d'une cible précise. Par exemple, dans le cas de l'analyse des émotions associées à un traitement particulier, les caractéristiques associées à cet objet médical sont bien connues (prix, tolérance, effets indésirables, *etc.*). Il est alors possible de présenter les associations entre cible et opinion comme dans le cas des avis de films ou de critiques (Hu & Liu, 2004; Zhuang *et al.*, 2006). Finalement, la principale limitation de cette contribution est de réduire l'identification de l'émotion seulement au cas où un mot porteur d'émotion(s) est présent. Comme souligné par (Osherenko & André, 2007), dans la plupart des cas, les personnes expriment leur émotions implicitement sans utiliser ces mots. Une émotion correspond à ce qu'une personne ressent à propos d'un fait et non au sentiment que la personne exprime à propos de ce fait. S'il est courant d'exprimer des sentiments sur les choses, il est plus fréquent de ressentir des émotions sans les exprimer explicitement. Dans des prochaines actions, nous prendrons cela en compte et essaierons d'identifier des émotions au-delà des cas explicites, car il est possible que dans les cas implicites l'identification des cibles diffère.

Références

BAKER C. F., FILLMORE C. J. & LOWE J. B. (1998). The Berkeley FrameNet project. In *Proceedings of COLING/ACL*, p. 86–90.

BETHARD S., YU H., THORNTON A., HATZIVASSILOGLOU V. & JURAFSKY D. (2004). Automatic extraction of opinion propositions and their holders. In J. G. SHANAHAN, J. WIEBE & Y. QU, Eds., *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text : Theories and Applications*, p. 22–24, Stanford, US.

BOAS H. C. (2002). Bilingual FrameNet dictionaries for machine translation. In M. G. RODRÍGUEZ & C. P. S. ARAUJO, Eds., *Proceedings of the Third International Conference on Language Resources and Evaluation*, volume IV, p. 1364–1371, Las Palmas.

CERINI S., COMPAGNONI V., DEMONTIS A., FORMENTELLI M. & GANDINI G. (2007). In *Language resources and linguistic theory : Typology, second language acquisition, English linguistics.*, Milano, IT : Franco Angeli Editore.

CHOI Y., CARDIE C., RILOFF E. & PATWARDHAN S. (2005). Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*.

DAS D., SCHNEIDER N., CHEN D. & SMITH N. A. (2010). Probabilistic frame-semantic parsing. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, p. 948–956, Stroudsburg, PA, USA : Association for Computational Linguistics.

DING X. & LIU B. (2007). The utility of linguistic rules in opinion mining. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, p. 811–812, New York, NY, USA : ACM.

DOWTY D. R., WALL R. E. & PETERS S. (1989). *Introduction to Montague Semantics*, volume 11. Dordrecht : D. Reidel.

EKMAN P. (1992). An argument for basic emotions. volume 6, p. 169–200.

HANCOCK J. T., TOMA C. L. & ELLISON N. B. (2007). The truth about lying in online dating profiles. In M. B. ROSSON & D. J. GILMORE, Eds., *CHI*, p. 449–452 : ACM.

HARB A., PLANTIÉ M., DRAY G., ROCHE M., TROUSSET F. & PONCELET P. (2008). Web Opinion Mining : How to extract opinions from blogs ? Categories and Subject Descriptors. In *International conference on Soft Computing as Transdisciplinary Science and Technology*, p. 211–217.

HU M. & LIU B. (2004). Mining opinion features in customer reviews. In *Proceedings of the 19th National Conference on Artifical Intelligence*, AAAI'04, p. 755–760 : AAAI Press.

KIM S.-M. & HOVY E. (2004). Determining the sentiment of opinions. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, p. 1367–1373.

KIM S.-M. & HOVY E. (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, SST '06, p. 1–8, Stroudsburg, PA, USA : Association for Computational Linguistics.

KOBAYASHI N., INUI K. & MATSUMOTO Y. (2007). Extracting aspect-evaluation and aspect-of relations in opinion mining. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 1065–1074.

KOZAREVA Z., NAVARRO B., VAZQUEZ S. & MONTOYO A. (2007). UA-ZBSA : a headline emotion classification through web information. In *4th International Workshop on Semantic Evaluations*, p. 334–337, Stroudsburg, PA, USA : ACL.

MELZI S., ABDAOUI A., AZÉ J., BRINGAY S., PONCELET P. & GALTIER F. (2014). Que ressentent les patients ? In *Proceedings of EGC'14*, p. 449–454.

MOHAMMAD S. M. & TURNEY P. D. (2010). Emotions Evoked by Common Words and Phrases :

Using Mechanical Turk to Create an Emotion Lexicon. In *Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, p. 26–34, Stroudsburg, PA, USA : ACL.

MUDINAS A., ZHANG D. & LEVENE M. (2012). Combining lexicon and learning based approaches for concept-level sentiment analysis. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '12, p. 5 :1–5 :8, New York, NY, USA : ACM.

NARAYANAN S. & HARABAGIU S. (2004). Question answering based on semantic structures. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA : Association for Computational Linguistics.

NEVIAROUSKAYA A., PRENDINGER H. & ISHIZUKA M. (2011). Affect analysis model : Novel rule-based approach to affect sensing from text. volume 17, p. 95–135, New York, NY, USA : Cambridge University Press.

OSHERENKO A. & ANDRÉ E. (2007). Lexical affect sensing : Are affect dictionaries necessary to analyze affect ? In *Proceedings of the 2Nd International Conference on Affective Computing and Intelligent Interaction*, ACII '07, p. 230–241, Berlin, Heidelberg : Springer-Verlag.

PANG B., LEE L. & VAITHYANATHAN S. (2002). Thumbs up ? : Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, p. 79–86, Stroudsburg, PA, USA : Association for Computational Linguistics.

POPESCU A.-M. & ETZIONI O. (2005). Extracting product features and opinions from reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, p. 339–346 : Computational Linguistics Association.

QUIRK R. (1985). *A Comprehensive grammar of the english language*. London [etc.] : Longman.

SIEGRIST J. (1994). *Emotions and Health in Occupational Life : New Scientific Findings and Policy Implications : Inauguration Speech Belle Van Zuylen Professorship*. Universiteit Utrecht.

STONE P. J. & HUNT E. B. (1963). *A Computer Approach to Content Analysis : Studies Using the General Inquirer System*. AFIPS '63 (Spring). New York, NY, USA : ACM.

STRAPPARAVA C. & MIHALCEA R. (2008a). Learning to identify emotions in text. In *Symposium on Applied Computing*, p. 1556–1560, New York, NY, USA : ACM.

STRAPPARAVA C. & MIHALCEA R. (2008b). Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC '08, p. 1556–1560, New York, NY, USA : ACM.

TAUSCZIK Y. R. & PENNEBAKER J. W. (2010). The psychological meaning of words : Liwc and computerized text analysis methods. volume 29, p. 24–54.

WHISSELL C. (1989). *The dictionary of affect in language*. Academic Press.

WIEBE J. (2000). Learning subjective adjectives from corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, p. 735–740 : AAAI Press.

WILSON T., WIEBE J. & HOFFMANN P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*, p. 347–354.

WU Y., ZHANG Q., HUANG X. & WU L. (2009). Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 3 - Volume 3*, p. 1533–1541, Stroudsburg, PA, USA : Association for Computational Linguistics.

ZHUANG L., JING F. & ZHU X.-Y. (2006). Movie review mining and summarization. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, p. 43–50, New York, NY, USA : ACM.

Construction et évolution de connaissances par confrontation de points de vue : prototype pour la recherche d'information scientifique

Guillaume Surroca[1], Philippe Lemoisson[1,2],
Clément Jonquet[1], Stefano A. Cerri[1]

[1] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM)
Université Montpellier 2 & CNRS
prenom.nom@lirmm.fr
[2] UMR Territoires, Environnement, Télédétection et Information Spatiale, CIRAD
Montpellier, France
philippe.lemoisson@cirad.fr

Résumé : Avec le Web 2.0, les utilisateurs, devenus contributeurs, ont pris une place centrale dans les processus de consommation et de production de connaissances ; cependant la paternité des contributions est souvent perdue lors de l'indexation de l'information. VIEWPOINTS est un formalisme de représentation des connaissances centré sur le point de vue individuel, humain ou artificiel. Nous considérons trois types d'objets de connaissance : les *documents* (supports), les *agents* (émetteurs) et les *topics* (descripteurs). Un *viewpoint* émis par un *agent* exprime son opinion sur la proximité entre deux objets. Les *viewpoints* permettent de définir et de calculer une distance entre objets qui évolue au fil des interactions (requêtes et retours d'utilisation) et de l'ajout de nouveaux *viewpoints*. Un prototype de moteur de recherche pour des données de publications scientifiques tirées de HAL-LIRMM montre comment VIEWPOINTS peut faire émerger, de façon transparente, une intelligence collective à partir des interactions des utilisateurs contributeurs.

Mots-clés : représentation des connaissances, construction collaborative de connaissances, indexation et recherche d'information, découverte interactive de connaissances, sérendipité, ingénierie des connaissances centrée utilisateurs, distance sémantique, Web 2.0.

1 Introduction

Le Web d'aujourd'hui est un espace de profusion de l'information où les échanges informels des réseaux sociaux cohabitent avec des jeux de données structurées pouvant alimenter des traitements automatiques (confrontation, intégration, raisonnement). C'est d'abord un espace où interagissent les humains, notamment depuis l'avènement du Web 2.0, à la fois en tant que producteurs/consommateurs de ressources, mais aussi en tant que commentateurs sur ces ressources. C'est aussi un espace peuplé d'agents artificiels invisibles, en charge notamment de tâches de fouille, de présentation, de traduction, d'indexation, etc. C'est donc un espace de partage de connaissances, où collaborent des agents humains et des agents artificiels, avec des connaissances collectives stabilisées dans les schémas du Web sémantique (vocabulaires, ontologies) [4] et des contributions individuelles spontanées du Web 2,0 (réseaux sociaux, recommandations) [16]. Face à cette complexité, les internautes sont confrontés à trois questions récurrentes :

- Comment trouver des documents ou des données dignes de confiance sur un sujet particulier ?

- Comment trouver les bonnes personnes pour échanger, argumenter et capitaliser sur un sujet particulier ?
- Peut-on faciliter les processus d'agrégation qui feront émerger de nouvelles connaissances au sein de la communauté ?

Partant de ce constat, notre travail a consisté à : (i) chercher un formalisme de représentation des connaissances qui puisse capturer les contributions individuelles tout en tissant des liens avec les schémas du Web sémantique ; (ii) chercher à proposer des protocoles régissant de façon transparente la recherche d'information. Ce document présente l'approche VIEWPOINTS et son formalisme, ainsi qu'un prototype permettant d'expérimenter dans le contexte d'une communauté réelle (les chercheurs du LIRMM) avec son corpus de connaissances (leurs publications scientifiques).

VIEWPOINTS permet à des *agents*[1] humains ou artificiels d'éliciter leurs points de vue (*viewpoints*) concernant la proximité sémantique entre des *objets* de connaissance du Web, que ce soit des supports (*documents*), des fournisseurs (*agents*), ou des descripteurs (*topics*) de la connaissance. L'approche VIEWPOINTS vise à tirer pleinement parti des points de vue explicites que des pairs, des collègues ou des compagnons ont précédemment exprimés. Un *viewpoint* est émis par un *agent* humain ou artificiel et porte sur la proximité sémantique entre deux objets de connaissance (*agents*, *documents* ou *topics*). Dans cette approche, la dynamique des points de vue individuels à travers les cycles réitérés de recherche d'information et de *feedback* fait émerger la connaissance collective sous forme de chemins renforcés ou atténués reliant les objets de connaissance au sein d'un graphe communautaire, selon la métaphore d'un cerveau dont les circuits neuronaux s'ajustent en continu au monde réel.

Le reste de l'article est organisé de la façon suivante : la section 2 pose le cadre de notre approche et présente notre vision. La section 3 présente le formalisme VIEWPOINTS. La section 4 illustre l'approche et aborde les questions mentionnées ci-dessus en mettant en œuvre un prototype dans un contexte d'utilisation réel : l'indexation des métadonnées d'une base de données bibliographique tirée de HAL-LIRMM. La section 5 discute les avantages et limitations de l'approche. Enfin, la section 6 tire les conclusions de cette première expérimentation pour orienter la poursuite de notre travail.

2 Etat de l'art et vision

L'espace de partage de connaissances où collaborent des agents humains et des agents artificiels peut être considéré comme un espace de consolidation-négociation de « points de vue », voire de « visions du monde » [20], [19], [8]. Dans les domaines où la consolidation est stabilisée apparaissent des ontologies contextuelles, cependant la majeure partie de l'espace reste peuplée de micro-expressions de sémantiques individuelles, avec parfois des éditeurs qui produisent un savoir consolidé en tissant des liens entre les visions du monde de plusieurs experts.

T. Gruber définit l'ontologie comme « la spécification d'une conceptualisation » [6]. Les ontologies sont le plus souvent constituées par un ensemble d'experts à l'issue d'un long processus de convergence de leurs représentations respectives (vision « par le haut »). Il est à noter que les ontologies contextuelles, même mises à jour quotidiennement (e.g., Gene Ontology), peuvent conduire à des situations où le rythme d'entretien du consensus est dépassé par le rythme d'évolution des connaissances de la communauté [15]. Quand le temps de la convergence est trop lent, l'étude des conditions d'émergence prend un intérêt accru ; c'est ainsi que certains travaux [1] sont dédiés à la possibilité de créer des ontologies « par le bas », directement à partir des interactions et de l'évolution d'un système. Toute la question réside alors dans la représentation de ces interactions. Dans la grande diversité des approches

[1] Les termes en italiques dans ce texte sont les termes réservés du formalisme ; ils sont définis dans la section 3.

qui existent aujourd'hui, nous distinguons deux grands courants : celui où l'agent porteur d'une sémantique est représenté explicitement dans le système, et celui où il ne l'est pas.

Pour commencer par le second, l'approche Topics Maps [19] [3] associe des thèmes (topics) à des ressources du Web via des occurrences et des contextes (scope), en mettant en relation des topics et des contextes ; cependant l'agent émetteur est ignoré. L'approche permet d'exploiter des relations topic-topic ou topic-document, ce qui est plus riche que la seule relation topic-document utilisée en recherche d'information. Le réseau de relations ainsi formé est très commode pour la navigation, mais reste difficilement exploitable de façon automatique (en termes d'algorithmes de graphe) de par la variété des types de relations et de leurs arités.

Dans le premier courant citons d'abord l'approche Hypertopic [20] où l'agent est introduit comme contributeur à une vue partielle portant sur une ou plusieurs ressources ; les visions du monde obtenues sont ainsi partagées entre plusieurs acteurs. Citons également Gesche et al. [5] qui propose une approche où les utilisateurs sont émetteurs et contradicteurs de points de vue, mais où aucun formalisme n'est proposé pour faciliter l'exploitation de l'espace de connaissances.

Les réseaux sociaux ou les sites collaboratifs sont bien évidemment les sources d'interactions principales pour ce premier courant. Ce sont des sources où puiser pour obtenir par émergence une sémantique collective à partir de l'expression des sémantiques individuelles. Le tagging social notamment est un mode d'interaction très répandu (e.g., flickr, delicious, last.fm) permettant à partir de micro-contributions d'obtenir et d'entretenir une folksonomie (de la contraction de folk et taxonomie). Les approches [15] [14] [18] considèrent de manière explicite les agents émetteurs de ces tags ; l'association d'un tag à une ressource par un agent étant généralement modélisée par un triplet agent-tag-document. Toutefois même si ces approches considèrent l'agent comme source explicite des associations, aucune ne considère l'agent comme objet potentiel de l'interaction c.-à-d. qu'il n'existe aucun triplet agent-agent-agent ou agent-agent-tag.

Un cas particulier intéressant est celui des systèmes contributifs, avec certaines approches utilisant le jeu comme levier de motivation, que ce soit pour l'enrichissement de folksonomie [9] ou pour la collecte d'informations lexicales [10] ; pour une revue des systèmes de recommandations qui mentionne en particulier les approches collaboratives voir [2]. Parmi les travaux récents qui visent à utiliser les tags d'une communauté pour enrichir un thésaurus, Limpens et al. [13] offrent aux utilisateurs de participer à la structuration d'un réseau de tags en mettant en place un protocole d'interaction impliquant les retours des utilisateurs et en recueillant et confrontant, grâce à des arbitres désignés, leurs points de vue sur les relations entre tags (related, broader, narrower, etc.).

Un élément de formalisation fédérateur du premier courant est la représentation basée sur le triplet *agent-document-topic*, où le terme *topic* englobe à la fois les tags cités ci-dessus et les concepts que l'on trouve dans les ontologies. Un autre élément fédérateur est le calcul de similarités entre ces objets (cf. [14] pour une revue et comparaison des différentes approches). En particulier, Quattrone et al. [17] offre une définition de la similarité basée sur le principe de renforcement mutuel : « deux tags sont considérés similaires s'ils ont été associés à des ressources similaires et vice-et-versa deux ressources sont considérées similaires si elles ont été associées à des tags similaires ». La prise en compte de similarités peut déboucher sur des mesures parfois appelées distances sémantiques [7][11] qui sont utiles pour l'annotation, la désambiguïsation, l'alignement de concepts ou la recherche d'information. Toutefois, si les distances sémantiques sont souvent utilisées pour déterminer la proximité entre *topics*, nous ne connaissons pas d'approche pour déterminer la proximité entre *topic*, *agent* et *document* d'une manière générique et symétrique qui permettrait d'exploiter pleinement le graphe qui sous-tend l'espace de partage de connaissances.

L'approche VIEWPOINTS se situe dans une vision qui cherche à exploiter directement la dynamique des interactions pour produire des connaissances par émergence, avec prise en compte explicite des agents, non seulement en tant que sources de points de vue, mais

également comme objets du graphe de connaissances. Notre but n'est pas de produire des ontologies mais plutôt de faire apparaître dans cet espace des lignes de force qui favorisent la sérendipité et la consolidation. Pour pouvoir à terme puiser dans les immenses volumes de données hétérogènes que sont les réseaux sociaux, nous nous appuyons sur un formalisme minimaliste centré sur une relation unique : le *viewpoint*, qui est défini dans la section 3.

Cette approche VIEWPOINTS s'inscrit dans une tentative de synthèse opérant les choix suivants:

- La représentation des connaissances s'appuie sur un hypergraphe constitué de triplets à la manière décrite dans [18], qui peut aussi être vu comme un graphe biparti avec d'un côté des objets et de l'autre des connecteurs. Dans le cas de VIEWPOINTS, les objets sont les *agents*, les *documents* et les *topics* ; nous traitons ces trois classes d'objets comme sous-classes d'une seule classe regroupant tous les objets de connaissance. Les connecteurs sont les *viewpoints* qui peuvent être des contributions initiales à la connaissance commune ou bien des *feedbacks* suite à une recherche d'information.
- L'accent est mis sur l'émergence au sein du graphe biparti, de même que [1].
- Le moteur évolutif du graphe est basé sur la dynamique « recherche d'information et *feedback* », à la manière de [13], mais sans procédure d'arbitrage.
- Nous définissons une distance métrique sur l'ensemble des objets de connaissance formé par les *agents*, les *documents* et les *topics* (alors que les distances sémantiques que l'on trouve dans la littérature s'appliquent à des sous-classes homogènes) et nous basons le calcul dynamique de cette distance sur l'ensemble des *viewpoints* (contributions initiales ou *feedbacks* des utilisateurs).

3 Le formalisme VIEWPOINTS

Nous présentons ci-dessous le formalisme VIEWPOINTS qui a déjà fait l'objet d'une description dans [12]. Considérons une collection O d'objets, comportant trois sous-classes A, D et T :

- les objets de la sous-classe A sont interprétés comme des *agents*. Les *agents* fournissent les point de vues ; ils sont soit humains (émetteurs de points de vue) soit artificiels (par exemple, les extracteurs de *topics*).
- les objets de la sous-classe D sont interprétés comme des *documents*. La notion de *document* unifie tous les supports de connaissance (textes, cartes, vidéos, etc.).
- les objets de la sous-classe T sont interprétés comme des *topics*. Le concept de *topic* unifie tous les moyens de description des *documents* ou des *agents* (mots clés, taxon, etc.), ou les thèmes de réflexion (sujets de discussion dans les fora).

Soit W l'ensemble de tous les couples constitués d'un *agent* de A et d'une paire d'objets quelconques de O ; les éléments de W sont de la forme $(a_i, \{o_j, o_k\})$ avec $a_i \in A$ et $o_j, o_k \in O$. En notant $a_i \rightarrow \{o_j, o_k\}$ l'élément $w = (a_i, \{o_j, o_k\})$, nous obtenons six formes de base : $a_1 \rightarrow \{d_1, t_1\}$, $a_1 \rightarrow \{t_1, t_2\}$, $a_1 \rightarrow \{d_1, d_2\}$, $a_1 \rightarrow \{a_2, t_1\}$, $a_1 \rightarrow \{a_2, d_1\}$ et $a_1 \rightarrow \{a_2, a_3\}$.

Un *viewpoint* est alors défini comme un triplet $v = (w, \alpha, \tau)$ où $w \in W$ et $\alpha, \tau \in \mathbb{R}$:

1. $w = a_1 \rightarrow \{o_2, o_3\}$ s'interprète de la façon suivante : « l'*agent* a_1 déclare une proximité entre les deux objets o_2 et o_3 ».
2. α est l'évaluation de cette proximité entre o_2 et o_3 telle qu'elle est donnée par l'*agent* a_1 à l'instant τ ; $\alpha \geq 0$.

Un exemple de *viewpoint* est l'association : « selon l'*agent* a_1, le *document* d_1 est pertinent relativement au *topic* t_1 à l'instant τ » ; ce *viewpoint* est formalisé en gardant des rôles symétriques pour d_1 et t_1 : $v = (a_1 \rightarrow \{d_1, t_1\}, +1, \tau)$.

Le *Knowledge Graph* (KG) est le graphe biparti suivant :

- les sommets de KG sont les éléments de $O \cup W$.
- les arêtes de KG sont obtenues à partir des éléments de W : chaque $w = a_1 \rightarrow \{o_2, o_3\}$ produit 3 arcs orientés : $a_1 \rightarrow w$, $w \rightarrow o_2$ and $w \rightarrow o_3$.
- les sommets pris dans W sont étiquetés par (α, τ) ; ce sont les *viewpoints.*

Les *viewpoints* de KG vont servir de base à la définition d'une distance sur O×O :

- soit $v = (w, \alpha, \tau) = (a_1 \rightarrow \{o_2, o_3\}, \alpha, \tau)$ un *viewpoint* connectant $\{o_2, o_3\}$, on écrit : $jump_v(\{o_2, o_3\}) = \alpha$ pour exprimer cette connexion élémentaire.[2]
- soit $W\{o_2, o_3\} = \{w \in W \mid \exists a_i \in A, (a_i, \{o_2, o_3\})\}$, l'ensemble des *viewpoints* connectant o_2 et o_3.
- l'ensemble des connexions entre deux objets $\{o_2, o_3\}$ dues aux différents *agents* constitue un « lien de proximité » nommé *synapse* dont on peut calculer la force en faisant la somme algébrique de tous les *jumps* reliant ces deux objets. On obtient une valeur positive:

$$synapse(\{o_2, o_3\}) = \sum_{W\{o_2, o_3\}} jump_v(\{o_2, o_3\})$$

- il est alors possible de considérer le graphe non orienté construit sur les sommets O reliés par les *synapses*, et de définir une distance métrique à partir des plus courts chemins dans ce graphe. Nous appelons *ψ-distance* cette distance métrique construite sur O×O.
- le *m-neighborhood* d'un objet 'o_q' de O est alors l'ensemble des objets 'o' ∈ O vérifiant : *ψ-distance* $(o_q, o) \leq m$. Le calcul du *m-neighborhood* s'inspire de l'algorithme du plus court chemin de Dijkstra en utilisant le paramètre m pour limiter la propagation dans le graphe. Ainsi, pour un objet 'o_q', *m-neighborhood*(o_q) renvoie les objets appartenant aux chemins partant de o_q et de longueur inférieure ou égale à m.

Dans KG, la dynamique des *viewpoints* est le reflet direct de la consolidation et de la confrontation des opinions individuelles au sein de la communauté. Toute recherche d'information s'appuie sur le calcul du *m-neighborhood*, lui-même basé sur l'ensemble des *viewpoints* présents au moment de la recherche. En retour, l'*agent* qui avait émis la requête est invité à évaluer la proximité entre l'objet initialement recherché et les objets du *m-voisinage* qui lui sont proposés. Ces « feedbacks » produisent des *viewpoints* nouveaux qui influenceront les prochains calculs de *ψ-distance*. Il y a donc coévolution des *synapses* au rythme des interactions communautaires.

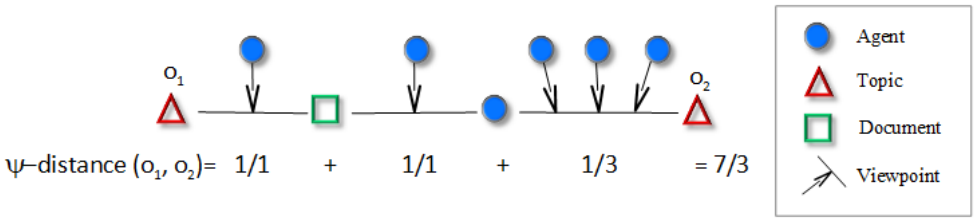


FIGURE 1 – Illustration du calcul de la distance entre deux objets de KG.

En outre, grâce au nouveau graphe construit à partir de O en évaluant toutes les *synapses* de KG, il devient possible de suivre l'évolution de trois structures émergentes : les réseaux de *documents* (bibliographies), les réseaux d'*agents* (sociogrammes), les réseaux de *topics* (onto-terminologies) comme illustré sur la figure 2. Cette analyse sera approfondie dans une prochaine publication.

[2] Dans la pratique, pour un instant τ donné, le graphe KG contiendra un seul *viewpoint* $(a_1 \rightarrow \{o_2, o_3\}, \alpha, \tau)$; la notation $jump_v$ est préférée à la notation $jump_a$ car l'agent a_1 peut émettre plusieurs *viewpoints* sur $\{o_2, o_3\}$ à des instants différents.

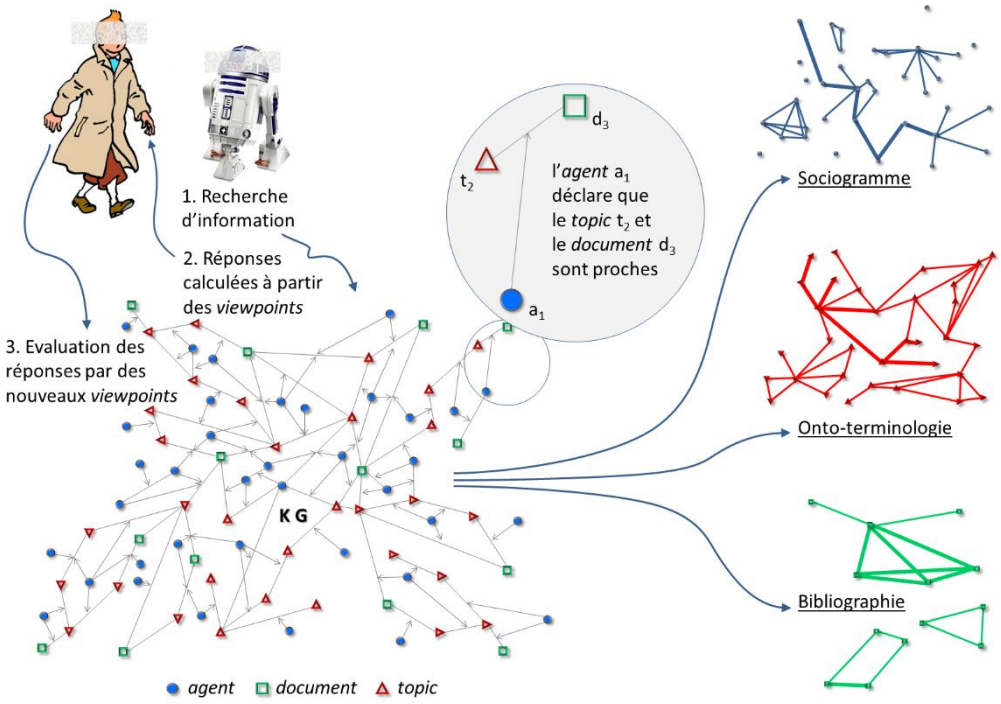


FIGURE 2 – Vue d'ensemble de l'approche VIEWPOINTS (toute ressemblance avec des *agents* existant ou ayant existé est purement fortuite). Par souci de lisibilité, α et τ ne sont pas représentés dans les *viewpoints*. Dans la partie droite, les traits représentent des synapses et leur épaisseur en représente la force.

4 Illustration de l'approche dans un contexte réel de recherche d'information

4.1 Ressource utilisée

Pour construire une application et tester cette approche sur des données réelles, nous avons choisi les ressources bibliographiques du HAL-LIRMM[3] comme corpus de données à indexer avec VIEWPOINTS. C'est une base de données de toutes les publications du LIRMM. Notre choix s'est porté vers cette ressource car :

- Chaque document est accompagné de métadonnées telles que les auteurs et les mots-clés choisis par ces auteurs pour décrire leurs publications. Cela en fait un jeu de données approprié pour illustrer le potentiel du formalisme.
- Nous avons accès à ce jeu de données de taille raisonnable.
- Ces données concernent nos collègues et notre laboratoire ce qui est donc pertinent pour évaluer/calibrer l'approche VIEWPOINTS et motiver nos collègues à fournir leur évaluation et leurs *viewpoints* lors du *feedback*.

4.2 Modèle adopté pour le graphe de connaissances

Dans un souci de comparaison et d'alignement avec le moteur de recherche fourni par HAL-LIRMM, nous avons sélectionné les métadonnées les plus simples pour l'initialisation du graphe de connaissance : pour un *document* d, pour chaque auteur a et pour chaque mot-clé t nous créons un *viewpoint* (a→{d, t}, +1, 0). Cette procédure est répétée sur chaque *document*. Dans notre application, basée sur les données de septembre 2013, 1663 *agents*, 5219 *documents* et 5846 *topics* nous donnent 42860 *viewpoints*.

[3] http://hal-lirmm.ccsd.cnrs.fr

Dans ce modèle, pour mettre en évidence l'impact des contributions, nous affectons un poids (α=3) aux viewpoints générés par les utilisateurs lors du *feedback* supérieur au poids des viewpoints créés lors de l'initialisation du graphe de connaissance avec les métadonnées.

Le prototype est implémenté en Java. Nous avons utilisé l'API d'affichage de graphe JUNG[4] afin d'avoir une visualisation du graphe de connaissance et des plus courts chemins.

4.3 Illustration de l'utilisation du prototype et du graphe de connaissance

La figure 3 illustre un sous-ensemble des objets au voisinage du *topic* 'Semantic Web'.[5] Une recherche de 'Semantic Web' sur HAL-LIRMM ne retourne que les objets qui sont directement liés à cette requête, c.-à-d. les articles ayant 'Semantic Web' dans leurs mots-clés ainsi que les auteurs de ces articles. Cependant, dans notre prototype, grâce à l'utilisation de la *ψ-distance*, la requête 'Semantic Web' renvoie en une seule recherche le *m-neighborhood* de ce *topic*, c.-à-d. tous les objets pour lesquels il existe un chemin de longueur inférieure ou égale à 'm' vers ce *topic*; par exemple 'Knowledge Management' ou 'Data Linking'.

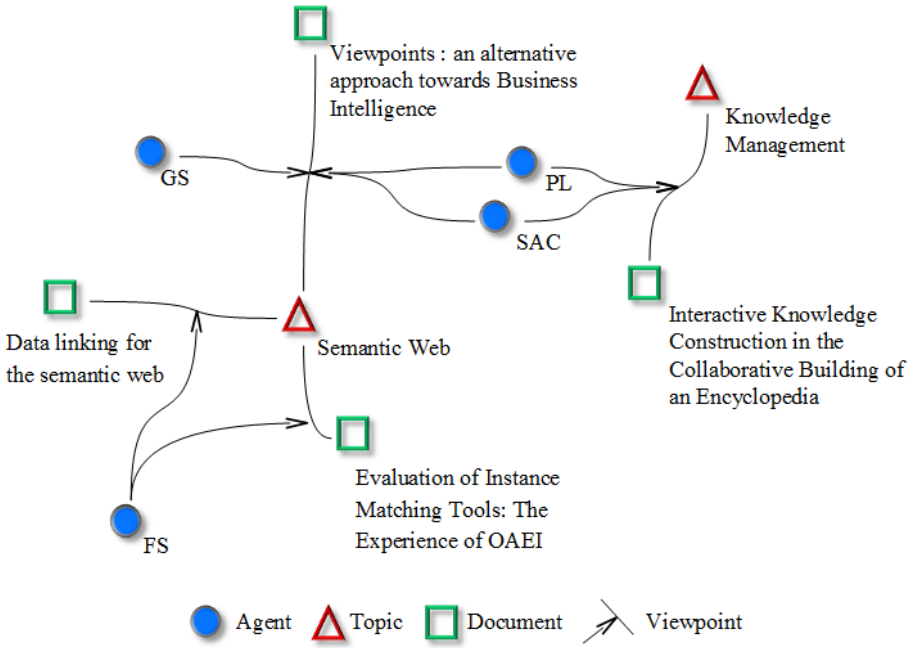


FIGURE 3 – Sous graphe d'objets au voisinage de 'Semantic Web' extraits à partir de KG.

Le cas traité pour illustrer l'approche se déroule en trois étapes :

Étape 1 : Guillaume Surroca (GS) exécute une recherche sur le *topic* 'Knowledge Management'.[6] L'interface présente alors les résultats dans trois onglets ('Documents', 'Agents', 'Topics') comme illustré dans la figure 4. L'*agent* Francois Scharffe (FS) y figure à une distance de 0,58 de l'objet recherché. Le prototype donne à l'utilisateur l'explication des résultats qu'il propose : en effet, l'utilisateur peut visualiser pour chaque résultat un des plus courts chemins reliant l'objet de la requête et le résultat dans le graphe de connaissances (bouton 'Path'). De plus, l'utilisateur peut valider chaque résultat (boutons 'Right') et émettre ainsi en guise de *feedback* de nouveaux *viewpoints* qui viendront nourrir le graphe pour les prochaines requêtes comme illustré dans la suite du scénario.

[4] Java Universal Network/Graph Framework, projet financé par la NSF américaine : http://jung.sourceforge.net

[5] Pour garder les schémas lisibles nous affichons dans le graphe de connaissances seulement les nœuds servant à l'illustration.

[6] Dans le prototype, un utilisateur saisit une chaine de caractères qui permet d'identifier l'objet de la requête (document, agent ou topic) par auto-complétion c'est-à-dire en se limitant explicitement aux objets de connaissance déjà présents dans le graphe.

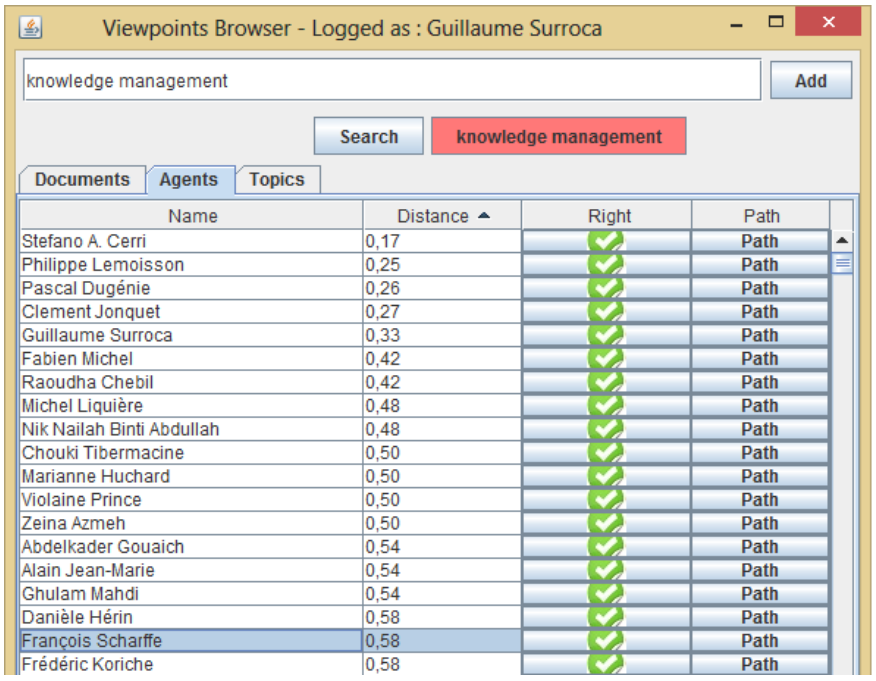


FIGURE 4 – Illustration d'une recherche sur 'Knowledge Management' dans l'interface. Le nom de l'utilisateur connecté apparait et permettra d'identifier l'émetteur des *viewpoints* lors du *feedback*.

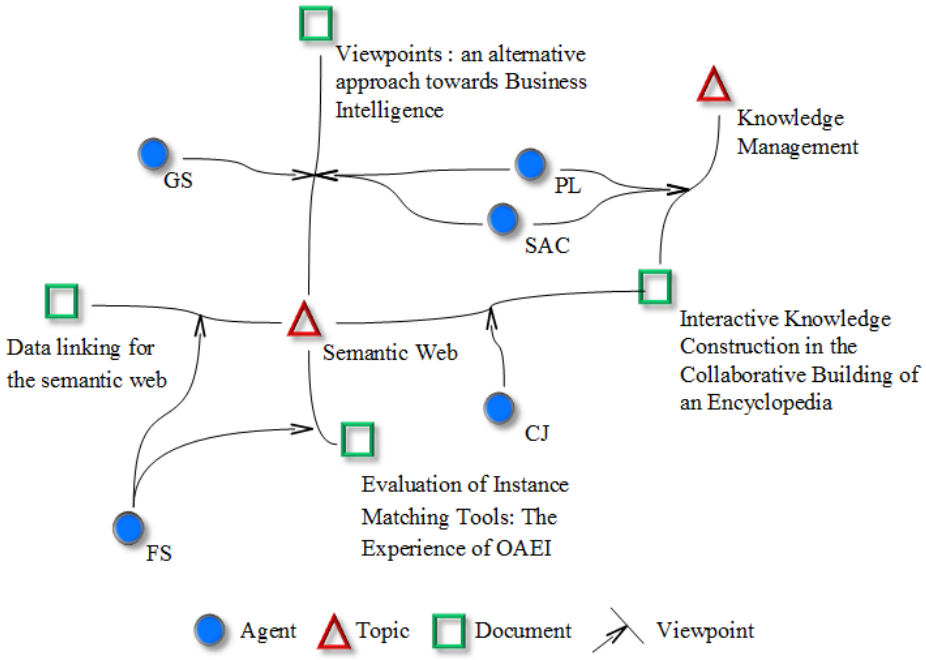


FIGURE 5 – Impact du *feedback* de CJ sur le graphe de connaissances.

Étape 2 : Ensuite, Clément Jonquet (CJ) fait une recherche sur le *topic* 'Semantic Web', et obtient comme résultat l'article 'Interactive Knowledge Construction in the Collaborative Building of an Encyclopedia' (IKC) ; son *feedback* consiste à approuver le résultat en émettant un nouveau *viewpoint* reliant ce *document* au *topic* 'Semantic Web'. La figure 5 illustre le graphe de connaissances après la contribution de CJ. Ce nouveau *viewpoint*, de poids $\alpha=3$ contribue à une *synapse* (IKC, Semantic Web) plus forte que les *synapses*

précédentes (SAC, Semantic Web) ou (PL, Semantic Web) ; ainsi il existe un nouveau plus court chemin et la distance diminue.

Étape 3 : Finalement, GS refait une recherche (figure 6) sur 'Knowledge Management', et cette fois-ci l'*agent* FS apparaîtra plus haut dans la liste des résultats (ordonnés par distances) étant donné que ces deux objets se sont rapprochés (0.50 < 0.58).

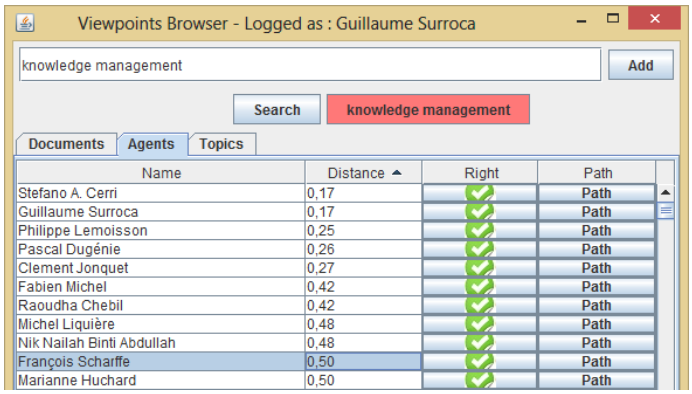


FIGURE 6 – Impact du *feedback* de CJ sur la recherche.

VIEWPOINTS offre un potentiel accru pour la recherche d'information :

- Pour un *agent* donné, le prototype retourne les *agents* au voisinage permettant d'identifier d'autres contributeurs ou collaborateurs potentiels. Il permet en outre d'identifier les *documents* proches de cet *agent* (sans se limiter aux publications dont il est auteur) et ses *topics* d'intérêts explicites (mot clés de ses publications) ou implicites (mots clés d'autres publications dont il est proche).
- Pour un *topic* donné, le prototype permet non seulement d'identifier les *documents* pertinents (comme n'importe quel moteur de recherche par mot clé) mais permet également d'identifier les experts pour ce *topic* et les *topics* proches dans le graphe de connaissance (illustration de la terminologie spécifique à une base de connaissances).
- Pour un *document* donné, un utilisateur peut trouver d'autres *documents* similaires (en plus des *topics* et *agents* proches).

5 Discussions

5.1 Aspects liés au corpus de connaissance et à la dynamique des contributions

Dans cette illustration de notre approche, le graphe de connaissances est obtenu par extraction d'une unique source de données ; il sera intéressant de l'enrichir par d'autres informations telles que les projets de recherche du LIRMM, l'organisation structurelle du laboratoire (équipes, membres et thématiques), etc. Ces autres jeux de données, une fois traduits en *viewpoints*, contribueront aux distances entre objets et participeront à la structuration du graphe de connaissances.

Par ailleurs, il faut noter que le graphe présenté est le graphe initial obtenu par extraction des métadonnées (à l'exception du *viewpoint* rajouté par CJ). C'est seulement après un nombre de requêtes et *feedbacks* suffisant que des lignes de force matérialisées par les synapses entre *viewpoints* émergeront : les *feedbacks* créent ou renforcent les synapses dans le graphe de connaissance. Par l'intermédiaire des calculs de voisinages, les *feedbacks* impactent donc directement les futures recherches d'autres utilisateurs, ce qui permet de parler d'intelligence collective.

Cette discussion amène la prise de conscience d'un challenge majeur dans notre approche : en utilisant les point de vues des utilisateurs comme contributions et sources de l'indexation

des objets de connaissance le problème de l'indexation est transféré vers d'autres problème : (i) comment motiver les contributions et (ii) comment tirer le maximum de ces contributions ?

5.2 Aspects liés aux choix de modélisation (façon dont nous avons transcrit le corpus)

Transformer des métadonnées en *viewpoints* suppose un ensemble de choix de modélisation. Ainsi, le choix exposé section 4.2 : « dans un *document* d, pour chaque auteur a et pour chaque mot-clé t nous créons un *viewpoint* (a→{d, t}, +1, 0) » est celui d'un modèle simple et immédiat. Il aurait été possible de rajouter par exemple des *viewpoints* exprimant de manière accentuée la paternité des *documents* (a→{a, d}, +10, 0), de façon à mettre en œuvre un modèle plus expressif. La détermination d'un modèle ayant un bon rapport expressivité/simplicité et son calibrage sont les prochaines étapes de notre travail sur ce prototype au moyen de poids différents et de typage des *viewpoints*.

En outre, étant donné que les *topics* sont pour le moment des mots-clés librement choisis par les auteurs lorsqu'ils ont enregistré leurs publications sur HAL-LIRMM le graphe de connaissance connaît les problèmes classiques des folksonomies (ambiguïté, polysémie, multilinguisme, etc.). Par exemple, certains *topics* peuvent être ambigus et créer de faux chemins ; ils peuvent également représenter le même concept. Une stratégie de remédiation serait l'intégration d'un *agent* artificiel exploitant une ontologie suffisamment riche pour éliminer les ambiguïtés et s'appuyant sur celle-ci pour exprimer sous forme de *viewpoints* des proximités sémantiques précises. Il est à noter que ce problème lié aux folksonomies n'existe pas avec des sources telles que PubMed (publications biomédicales) respectant une terminologie standardisée (MeSH).

Cette discussion amène des questions plus générales : comment extraire sous forme de *viewpoints* des données ou métadonnées explicites formalisées dans des bases de connaissances, des jeux de données, ou des ontologies ? Comment extraire des *viewpoints* quand ils sont implicites (e.g., exprimés sous forme textuelle) ? Comment lier nos *topics* aux schémas du Web sémantique (vocabulaires et ontologies existantes) ?

5.3 Aspects liés au formalisme lui même

Le fait d'obtenir dans les résultats d'une recherche des objets qui ne sont pas directement liés (par les métadonnées) à l'objet de la requête est la principale source de sérendipité. Il est dans la nature de la *ψ-distance* d'ouvrir des chemins par transitivité : ainsi, dans le prototype, deux *topics* attachés au même *document* deviennent automatiquement « un peu proches », ce qui peut ne pas toujours être approprié. Il ne faut pas oublier cependant que c'est la dynamique de l'interaction qui produit les liens sémantiques les plus forts : une proximité discutable doit pouvoir être remise en question par des *viewpoints*. La réflexion concernant une gestion discriminante des chemins hétérogènes (*topic-document-topic* par exemple) par l'algorithme de plus court chemin est en cours.

5.4 La question de l'évaluation

Que ce soit pour perfectionner le formalisme, pour choisir et calibrer un modèle, ou pour établir la preuve de concept dans un scénario réel, la question de l'évaluation du gain en intelligence collective est importante et difficile. Il s'agit en effet d'évaluer un apprentissage collectif, sans objectifs initiaux. Nous avons des pistes pour cet aspect : (i) l'observation de l'évolution de la structure du graphe de connaissances au fil des interactions ; (ii) l'observation qualitative du flux de *viewpoints*. La réflexion est en cours pour trouver des benchmarks et un protocole de simulation satisfaisants. Des benchmarks de recherche d'information nous permettront d'évaluer (en termes de précision/rappel) le scénario de recherche *topic-document*. Nous prévoyons également une évaluation en situation réelle par les utilisateurs.

6 Conclusions et perspectives

En traitant le corpus des publications des chercheurs du LIRMM, nous avons montré l'opérationnalité de l'approche VIEWPOINTS dans un contexte de recherche d'information, et nous avons apporté des éléments de réponse aux questions énoncées dans l'introduction :

- Le graphe de connaissances réifie en toute transparence la sémantique collective de la communauté. Il contient toutes les explications concernant l'émergence de cette sémantique à partir des contributions. Le processus de réponse aux requêtes est donc lui aussi transparent, permettant à l'utilisateur de trouver des documents ou des données dignes de confiance sur ses sujets d'intérêt.
- La géographie de la connaissance ainsi produite permet aisément de trouver les bonnes personnes pour échanger, argumenter et capitaliser sur un sujet particulier.
- Au fil des interactions, la structure du graphe élicite une proximité sémantique entre certains *topics*, en reflétant les points de vue des membres de la communauté. Ceci est un premier pas vers des processus d'agrégation susceptibles de faire émerger de nouvelles connaissances.

Ce premier prototype nous a permis d'expérimenter le calcul de la *ψ-distance* sur des données réelles, mais surtout de valider l'aptitude du formalisme à supporter un modèle traitant la recherche d'information scientifique. Pour achever la preuve de concept, nous envisageons deux scénarios d'expérimentation à plus grande échelle afin de tenter d'appréhender le gain en intelligence collective :

- Un scénario centré sur les ontologies et les ressources de données biomédicales où l'objectif sera d'intégrer sous forme de *viewpoints* : (i) des annotations (manuelles ou automatiques) basées sur les ontologies ou (ii) des données structurées comme des données liées sur le Web. L'objectif sera de montrer les apports de l'approche pour l'indexation sémantique.
- Un scénario centré sur une thématique environnementale, où l'objectif sera d'exploiter les connaissances explicites et implicites des chercheurs du Cirad pour faire émerger une connaissance communautaire dans une dynamique favorisant la controverse.

La question de l'évaluation évoquée section 5 sera abordée simultanément à la définition de ces scénarios. Ces scénarios nous permettront de poursuivre les travaux sur le formalisme et sur l'exploitation des graphes de connaissances en bénéficiant des retours des utilisateurs. Nous envisageons l'exploitation de la dimension temporelle, simultanément avec des études de clusterisation, pour observer l'évolution des graphes et étudier les dynamiques d'évolution des connaissances au sein des communautés.

Remerciements

Ce travail a bénéficié des soutiens du Cirad et du projet SIFR (Semantic Indexing of French Biomedical Resources) financé en partie par le programme JCJC de l'Agence nationale de la Recherche (ANR-12-JS02-01001), l'Université Montpellier 2, le CNRS et l'Institut de Biologie Computationnelle de Montpellier.

Références

[1] ABERER, K., CUDRE-MAUROUX, P., OUKSEL, A., CATARCI, T., HACID, M.-S., ILLARRAMENDI, A., KASHYAP, V., MECELLA, M., MENA, E., NEUHOLD, E., TROYER, O., RISSE, T., SCANNAPIECO, M., SALTOR, F., SANTIS, L., SPACCAPIETRA, S., STAAB, S., AND STUDER, R. Emergent Semantics Principles and Issues. In *Database Systems for Advanced Applications*, Y. Lee, J. Li, K.-Y. Whang, and D. Lee, Eds., vol. 2973 of *Lecture Notes in Computer Science*. Springer, 2004, pp. 25–38.

[2] ADOMAVICIUS, G., AND TUZHILIN, A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering 17*, 6 (June 2005), 734–749.

[3] CAUSSANEL, J., CAHIER, J.-P., ZACKLAD, M., AND CHARLET, J. Les Topic Maps sont-ils un bon candidat pour l'ingénierie du Web Sémantique ? In *13èmes journées francophones d'Ingénierie des Connaissances, IC'02* (Rouen, France, May 2002), p. 12.

[4] GANDON, F., FARON-ZUCKER, C., AND CORBY, O. *Le web sémantique - Comment lier les données et les schémas sur le web ?* Dunod, 2012.

[5] GESCHE, S., CAPLAT, G., AND CALABRETTO, S. Managing Difference of Opinion in Semantic Structures. In *International Workshop on Semantically Aware Document Processing and Indexing, SADPI'07* (Montpellier, France, May 2007), H. Betaille, J.-Y. Delort, M.-L. Mugnier, J. Nanard, and M. Nanard, Eds., ACM, pp. 79–86.

[6] GRUBER, T. R. A translation approach to portable ontologies. *Knowledge Acquisition 5*, 2 (June 1993), 199–220.

[7] HARISPE, S., RANWEZ, S., JANAQI, S., AND MONTMAIN, J. The Semantic Measures Library and Toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics* (October 2013).

[8] IACOVELLA, A., BENEL, A., PETARD, X., AND HELLY, B. *La redocumentarisation du monde*. Cépaduès, 2006, ch. Corpus scientifiques numérisés : Savoirs de référence et points de vue des experts, pp. 117–130.

[9] KRAUSE, M., AND ARAS, H. Playful tagging: folksonomy generation using online games. In *18th International Conference on World Wide Web, WWW'09* (Madrid, Spain, 2009), pp. 1207–1208.

[10] LAFOURCADE, M. Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *7th International Symposium on Natural Language Processing, SNLP'07* (Pattaya, Chonburi, Thailand, December 2007), p. 7.

[11] LEE, W.-N., SHAH, N. H., SUNDLASS, K., AND MUSEN, M. A. Comparison of Ontology-based Semantic-Similarity Measures. In *American Medical Informatics Association Annual Symposium, AMIA'08* (Washington DC, USA, November 2008), pp. 384–388.

[12] LEMOISSON, P., SURROCA, G., AND CERRI, S. A. Viewpoints: An Alternative Approach toward Business Intelligence. In *eChallenges e-2013 Conference* (Dublin, Irland, October 2013), p. 8.

[13] LIMPENS, F., GANDON, F., AND BUFFA, M. Un cycle de vie complet pour l'enrichissement sémantique des folksonomies. In *11ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances, EGC'11* (Brest, France, Janvier 2011), A. Khenchaf and P. Poncelet, Eds., vol. E-20 of *Revue des Nouvelles Technologies de l'Information*, Hermann, pp. 389–400.

[14] MARKINES, B., CATTUTO, C., MENCZER, F., BENZ, D., HOTHO, A., AND STUMME, G. Evaluating Similarity Measures for Emergent Semantics of Social Tagging. In *18th International Conference on World Wide Web, WWW'09* (Madrid, Spain, 2009), pp. 641–650.

[15] MIKA, P. Ontologies Are Us: A Unified Model of Social Networks and Semantics. In *4th International Semantic Web Conference, ISWC'05* (Galway, Ireland, November 2005), Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen, Eds., vol. 3729 of *Lecture Notes in Computer Science*, Springer, pp. 522–536.

[16] O'REILLY, T. What Is Web 2.0. Oreilly's blog post, September 2005.

[17] QUATTRONE, G., FERRARA, E., MEO, P. D., AND CAPRA, L. Measuring Similarity in Large-scale Folksonomies. In *23rd International Conference on Software Engineering and Knowledge Engineering, SEKE'11* (Miami Beach, FL, USA, July 2011), pp. 385–391.

[18] SPECIA, L., AND MOTTA, E. Integrating Folksonomies with the Semantic Web. In *4th European Semantic Web Conference, ESWC'07* (Innsbruck, Austria, June 2007), E. Franconi, M. Kifer, and W. May, Eds., vol. 4519 of *Lecture Notes in Computer Science*, Springer, pp. 624–639.

[19] STEVE PEPPER, G. O. G. Towards a General Theory of Scope. In *Extreme Markup Languages Conference* (Montreal, Canada, August 2001), p. 4.

[20] ZACKLAD, M., BENEL, A., ZAHER, L., LEJEUNE, C., CAHIER, J.-P., AND ZHOU, C. Hypertopic: une métasémiotique et un protocole pour le Web socio-sémantique. In *18ème journées francophones d'Ingénierie des Connaissances, IC'07* (Grenoble, France, July 2007), F. Trichet, Ed., Cépaduès, pp. 217–228.

Vers des recommandations plus personnalisées dans les folksonomies

Mohamed Nader Jelassi[1,2,3], Sadok Ben Yahia[1] et Engelbert Mephu Nguifo[2,3]

[1] Université Tunis El Manar. Faculté des Sciences de Tunis, Tunis, Tunisie.

[2] Clermont Université, Université Blaise Pascal, LIMOS, BP 10448, F-63000 Clermont-Ferrand, France.

[3] CNRS, UMR 6158, LIMOS, F-63171 Aubière, France.
{nader.jelassi@isima.fr, sadok.benyahia@fst.rnu.tn, engelbert.mephu_nguifo@univ-bpclermont.fr}

Résumé :
Plusieurs approches ont été proposées dans la littérature pour personnaliser les recommandations dans les *folksonomies*. Dans ce papier, nous considérons une nouvelle dimension dans les *folksonomies* comme information supplémentaire pour offrir aux utilisateurs une recommandation plus ciblée et mieux conforme à leurs besoins. Cela passe par un regroupement des utilisateurs ayant des intérêts communs sous forme de structures appelées quadri-concepts. Notre approche, dans laquelle nous répondons également au challenge de cold start, est ensuite évaluée sur deux jeux de données du monde réel, MOVIELENS et BOOKCROSSING. Cette évaluation comprend une mesure de la précision et du rappel, une évaluation sociale ainsi que plusieurs métriques d'évaluation comme la diversité, la couverture ou la scalabilité.

Mots-clés : Folksonomie, Personnalisation, Recommandation, Cold Start, Précision

1 Introduction et Motivations

Une *folksonomie* désigne un système de classification collaborative par les internautes (Mika (2007)). L'idée est de permettre à des utilisateurs de partager et de décrire des objets via des mots-clés (tags) librement choisis. Formellement, une *folksonomie* est composée de trois ensembles : un ensemble $\mathcal{U}$ d'utilisateurs, un ensemble $\mathcal{T}$ de tags (ou étiquettes) et un ensemble $\mathcal{R}$ de ressources (films, livres, sites web, photos, etc.). Les utilisateurs sont les acteurs principaux du système et contribuent au contenu par l'ajout de ressources et l'affectation de tags. Cependant, il s'avère que le choix de tags et de ressources partagées par un utilisateur d'une *folksonomie* varie selon plusieurs critères : le genre, l'âge ou encore la profession de celui qui partage l'information. Ainsi, les *folksonomies* doivent tenir compte de telles informations lors de la recommandation de tags ou de ressources. Cela a incité les chercheurs à proposer des systèmes de recommandation personnalisés afin de suggérer les tags et ressources les plus appropriés aux utilisateurs et de répondre aux besoins de chaque utilisateur. En effet, le domaine de personnalisation tente de fournir des solutions afin d'aider les utilisateurs à partager les bons tags et les bonnes ressources parmi le très grand nombre de données dans les *folksonomies*. Ainsi, un système de recommandation offre à l'utilisateur une liste de tags ou de ressources recommandés qu'il est susceptible d'aimer et lui permet de trouver plus facilement ses tags et ressources préférés dans la *folksonomie* (Ricci *et al.* (2011)). De plus, la personnalisation tente d'aider les utilisateurs à aborder le problème de surcharge d'information (Das *et al.* (2012)). Et pour réussir ou tenter de répondre au mieux aux attentes de chaque utilisateur de la *folksonomie*, il est utile d'avoir plus d'informations sur lui. En effet, son âge, sa profession ou sa localisation sont des informations qui sont susceptibles de nous aider dans le processus de personnalisation

de recommandation. Pour atteindre cet objectif, nous considérons une nouvelle dimension dans une *folksonomie*, classiquement composée de trois dimensions (utilisateurs, tags et ressources), et nous proposons une approche de regroupement des utilisateurs aux intérêts équivalents sous forme de structure appelées concepts quadratiques (Jelassi *et al.* (2013)). Cette quatrième dimension peut recouvrir différents aspects : par exemple le profil (genre, âge, profession, $ldots$) comme mentionné ci-dessus, ou le temps si on veut étudier la dynamique temporelle des *folksonomies*. Dans ce papier, nous traitons la quatrième dimension de manière indifférente pour l'aspect méthodologique, mais afin de comparer notre méthode avec les travaux de la littérature, nous focaliserons sur l'aspect profil.

Une question se pose alors : pourquoi les concepts quadratiques ? D'un côté, si on peut facilement étudier les tags utilisés par un seul utilisateur sur une ressource, il est évident de constater que la tâche devient rapidement intraitable lorsque cela implique plusieurs utilisateurs et plusieurs ressources. D'un autre côté, les tags (ou ressources) recommandés s'avèrent ne pas être très spécifiques (Jäschke *et al.* (2007)), *i.e.,* des tags qui sont des mots "bateau" ou bien des ressources vagues ne correspondant pas aux besoins spécifiques de l'utilisateur. Grâce aux concepts quadratiques, nous pouvons résoudre ces problèmes. En effet, d'un côté, les concepts quadratiques sont des strucutres regroupant les tags et ressources en commun à un ensemble maximal d'utilisateurs. D'un autre côté, dans un concept quadratique, les tags et ressources qui ont été utilisés en combinaison seront regroupés d'où un résultat plus spécifique et répondant au besoin de notre système de recommandation. Ces concepts sont une représentation réduite de la *folksonomie* qui peut contenir des milliers de quadruplets dans la vraie vie. Une fois extraits, ces quadri-concepts sont utilisés pour notre algorithme de recommandation personnalisée multi-mode (utilisateurs, tags et ressources). Nous menons ensuite une évaluation étendue de notre système de recommandation sur deux jeux de données : tout d'abord, nous calculons la précision et le rappel de notre système, ensuite nous analysons ses différentes propriétés comme la diversité ou la scalabilité et nous menons une évaluation sociale sur différents utilisateurs. De plus, nous étudions la capacité de notre système à proposer des recommandations aux nouveaux utilisateurs, *i.e.,* le problème de cold start (Lam *et al.* (2008)).

Le reste du papier est organisé comme suit : nous étudions les principales approches de la littérature dans la Section 2. Dans la Section 3, nous proposons un système personnalisé de recommandation. Nous menons une étude expérimentale dans la Section 4. Enfin, nous concluons notre papier avec des perspectives pour nos travaux futurs dans la Section 5.

2 Travaux connexes

Dans un souci d'améliorer les recommandations dans les *folksonomies*, plusieurs travaux ont été proposés dans la littérature. Dans (Diederich & Iofciu (2006)), les auteurs utilisent la *"personomie"* d'un utilisateur, *i.e.,* les tags qui lui sont relatifs, afin de lui recommander des utilisateurs ayant partagé des tags et ressources similaires. Tout d'abord, ils construisent un profil pour chaque utilisateur. Ensuite, à partir de ce profil, les auteurs sont capables de recommander des utilisateurs (dits *collaborateurs*) en utilisant une mesure de similarité entre utilisateurs. Cette mesure, qui s'appuie uniquement sur les tags utilisés par les utilisateurs, n'offre pas une information complète sur les utilisateurs. Plus récemment, dans (Hu *et al.* (2011)), les auteurs se basent à la fois sur l'historique de tagging (tags et ressources) des utilisateurs et sur leurs contacts sociaux. La limite de cette approche est qu'elle requiert qu'un utilisateur doit posséder

des contacts sociaux afin d'avoir des recommandations de tags. Dans (Jäschke *et al.* (2007)), Hotho *et al.* ont proposé des recommandations de tags dans les *folksonomies* basées sur les tags les plus utilisés. Cependant, ces recommandations ne sont absolument pas personnalisées étant donné que les mêmes tags sont proposés à chaque utilisateur. Lipczak a proposé dans (Lipczak (2008)) un système de recommandation de tags en trois étapes. À partir des tags annotés aux ressources, l'auteur ajoute des tags proposés par un lexique basé sur les co-occurrences de tags sur les mêmes ressources. Ensuite, le système filtre les tags déjà utilisés par l'utilisateur. Toutefois, malgré cette étape de filtrage, la recommandation ne paraît pas être personnalisée étant donné qu'elle cherche des tags co-occurrant sur d'autres annotations. L'approche revient ensuite à enlever les tags précédemment annotés par l'utilisateur de ceux qui sont suggérés. Dans (Landia & Anand (2009)), les auteurs ont proposé une nouvelle approche combinant la similarité à la fois entre ressources et entre utilisateurs afin de recommander des tags personnalisés. En effet, deux utilisateurs sont considérés comme similaires s'ils ont assigné les mêmes tags aux mêmes ressources. Toutefois, il est rare de trouver pareille situation dans des *folksonomies* où les tags utilisés par deux utilisateurs sur les mêmes ressources sont identiques.

Dans notre approche, nous insistons sur le nécessaire recours à des informations supplémentaires et à les combiner à l'historique de tagging afin d'améliorer les recommandations. Toutes ces informations seront représentées par des quadri-concepts. Ainsi, dans ces structures, nous nous focalisons non seulement sur les tags/ressources les plus utilisés, mais également sur ceux qui ont été utilisés en combinaison, obtenant ainsi un résultat plus spécifique. Contrairement aux approches de la littératures qui se limitent à l'information <utilisateur, tag, ressource>, nous étendons ce triplet par l'information contenue dans la quatrième dimension.

3 Un nouveau système de recommandation personnalisée

Nous commençons par présenter une extension de la notion de *folksonomie* (Jäschke *et al.* (2008)) par l'ajout d'une quatrième dimension.

Une **v-folksonomie** est un ensemble de tuples $\mathcal{F}_v = (\mathcal{U}, \mathcal{T}, \mathcal{R}, \mathcal{V}, \mathrm{Y})$ où $\mathcal{U}$, $\mathcal{T}$, $\mathcal{R}$ et $\mathcal{V}$ sont des ensembles finis dont les éléments sont appelés **utilisateurs**, **tags**, **ressources** et **variables**. $\mathrm{Y} \subseteq \mathcal{U} \times \mathcal{T} \times \mathcal{R} \times \mathcal{V}$ représente une relation quadratique où chaque élément $y \subseteq \mathrm{Y}$ peut être représenté par un quadruplet : $y = \{(u, t, r, v) \mid u \in \mathcal{U}, t \in \mathcal{T}, r \in \mathcal{R}, v \in \mathcal{V}\}$ ce qui veut dire que l'utilisateur u a annoté la ressource r via le tag t à travers la variable v. Nous considérons que deux utilisateurs sont *proches* s'ils partagent au moins une même variable en commun.

Nous définissons maintenant un concept quadratique (Jelassi *et al.* (2013)).

Un **concept quadratique** (ou quadri-concept) d'une *v-folksonomie* $\mathcal{F}_v = (\mathcal{U}, \mathcal{T}, \mathcal{R}, \mathcal{V}, \mathrm{Y})$ est un quadruplet (U, T, R, V) avec $U \subseteq \mathcal{U}$, $T \subseteq \mathcal{T}$, $R \subseteq \mathcal{R}$ et $V \subseteq \mathcal{V}$ avec $U \times T \times R \times V \subseteq \mathrm{Y}$ tel que le quadruplet (U, T, R, V) est maximal. Un quadri-concept est donc la version quadri-dimensionnelle d'un ensemble fermé. Par ailleurs, la quatrième dimension $\mathcal{V}$ peut recouvrir différents aspects (*e.g.,* le profil, le temps). Dans ce papier, nous traitons la quatrième dimension de manière indifférente pour l'aspect méthodologique, mais afin de comparer notre méthode avec les travaux de la littérature, nous focaliserons sur l'aspect profil.

Afin de permettre l'extraction de l'ensemble de quadri-concepts fréquents à partir d'une *v-folksonomie* donnée, nous pouvons utiliser l'un des deux algorithmes de la littérature dédiés à cette tâche : QUADRICONS (Jelassi *et al.* (2013)) ou DATAPEELER (Cerf *et al.* (2009)). Les deux algorithmes prennent en entrée une *v-folksonomie* ainsi que quatre seuils minimaux de

support (un pour chaque dimension) et donnent en sortie l'ensemble de **quadri-concepts** vérifiant ces seuils. Un quadri-concept **fréquent** est un quadri-concept dont chaque ensemble (utilisateur, tag, ressource et variable) a une cardinalité supérieure ou égale à son seuil de support correspondant. Par ailleurs, il est important de noter que, même si l'étape d'extraction des quadri-concepts est une phase qui peut avoir une complexité exponentielle, elle se passe hors-ligne et n'est exécutée qu'une seule fois. En effet, notre algorithme PERSOREC parcourt des quadri-concepts déjà extraits. Ainsi, notre système de recommandation ne souffre pas du coût d'extraction des quadri-concepts à chaque recommandation. Cette phase peut donc être vue comme un pré-traitement à la phase de recommandation. Dans ce qui suit, nous introduisons notre algorithme de recommandation personnalisée pour les *folksonomies* (Algorithme 1) qui donne en sortie trois différents ensembles : un ensemble d'utilisateurs proposés, un ensemble de tags suggérés et un ensemble de ressources recommandées.

Algorithme 1 : PERSOREC

Données : l'ensemble des quadri-concepts fréquents $\mathcal{QC}$, un utilisateur u avec sa variable v et une ressource r.

Résultats : l'ensemble d'utilisateurs proposés $\mathcal{PU}$, l'ensemble des tags suggérés $\mathcal{ST}$ et l'ensemble des ressources recommandées $\mathcal{RR}$.

```
1  début
2      pour chaque quadri-concept qc ∈ QC faire
3          si v ∈ qc.Variables alors
4              si u ∉ qc.Utilisateurs alors
5                  /*Proposition d'utilisateurs */
6                  PU = PU ∪ qc.extent ;
7                  /*Suggestion de Tags*/
8                  si r ∈ qc.Ressources alors
9                      ST = ST ∪ qc.Tags ;
10                 /*Recommandation de Ressources */
11                 RR = RR ∪ qc.Ressources ;
12     retourner (PU,ST,RR) ;
13 fin
```

PERSOREC opère comme suit : dans la Ligne 3, il parcourt l'ensemble des quadri-concepts fréquents en cherchant ceux dont les utilisateurs sont proches de u selon la variable v. Le test de la Ligne 4 permet de filtrer les tags et ressources déjà partagés par l'utilisateur u ; cette stratégie est inspirée par celle de (Lipczak (2008)). Ensuite, pour chaque tâche, PERSOREC fonctionne comme suit : pour la tâche de *Proposition d'utilisateurs* (Ligne 6), c'est la partie *utilisateurs* du quadri-concept qc qui est ajoutée à l'ensemble $\mathcal{PU}$ des utilisateurs proposés. Cette tâche aide à connecter les utilisateurs qui ont des intérêts communs et aide également à promouvoir le partage de ressources. Pour la tâche de *Suggestion de tags* (Lignes 8 et 9), le but est de suggérer des tags personnalisés à un utilisateur qui souhaite ajouter une ressource à la *folksonomie*. Cette tâche a plusieurs avantages : elle rappelle à l'utilisateur ce dont une ressource s'agit, accroît l'annotation des ressources et permet de consolider le vocabulaire des utilisateurs

(Ricci *et al.* (2011)). Pour cette tâche, nous ajoutons donc les tags affectés à la ressource r par les utilisateurs proches de u à l'ensemble $\mathcal{ST}$. Quant à la tâche de *Recommandation de ressources* (Ligne 12), le but est de proposer une liste personnalisée de ressources conforme aux intérêts de l'utilisateur u ; ces ressources sont ajoutées à l'ensemble $\mathcal{RR}$. Dans ce qui suit, nous évaluons notre approche sur deux jeux de données standard pour la recommandation..

4 Résultats et Discussion

Les deux jeux de données du monde réel utilisés pour notre évaluations sont : tout d'abord, le jeu de données filmographique MOVIELENS (*http ://movielens.umn.edu/*) qui est un système de recommandation et un site web communautaire qui permet aux utilisateurs de partager des films en les annotant par des tags. Le jeu de données, utilisé pour nos expérimentations, est téléchargeable gratuitement (*http ://www.grouplens.org/node/73*) et contient 95580 tags appliqués à 10681 films par 71567 utilisateurs (*e.g., <Alex*, *X-Files*, *sciencefiction>*). Le second jeu de données utilisé est BOOKCROSSING (*http ://www.bookcrossing.com/*) qui est un "club de lecture" en ligne dont le but est de "faire du monde entier une bibliothèque". Contrairement à MOVIELENS, les utilisateurs de BOOKCROSSING n'utilisent pas les tags pour annoter les ressources, *i.e.,* les livres, mais plutôt sur les notes. Ainsi, les utilisateurs choisissent une note entre 1 et 10, *i.e.,* plus la note est élevée, plus l'appréciation est meilleure. Le jeu de données utilisé est téléchargeable gratuitement (*http ://www.grouplens.org/node/74*) et contient 278858 utilisateurs, 1149780 notes et 271379 livres (*e.g., <Regina*, *DaVinciCode*, *9>*). Pour les besoins de comparaison avec les approches de la littérature, nous avons choisi, dans ce qui suit, le profil des utilisateurs pour modéliser la variable v. Ainsi, nous considérons à présent que deux utilisateurs sont *proches* s'ils partagent au moins une même information de profil en commun (*e.g.,* un même âge, une même profession, etc.). À cet effet, des informations supplémentaires sur les utilisateurs sont disponibles dans les deux jeux de données et forment le profil des utilisateurs (la quatrième dimension d'une *p-folksonomie*) et qui renseigne, pour MOVIELENS, sur le **genre** de l'utilisateur (masculin ou féminin) et sa **profession** (au nombre de 21, qui peut être éducateur, écrivain, étudiant, scientifique, etc.). Pour BOOKCROSSING, nous avons des informations sur la localisation des utilisateurs. Par ailleurs, les deux jeux de données fournissent des informations sur l'**âge** des utilisateurs qui est divisé en cinq tranches : (*i*) $7-18$ ans ; (*ii*) $19-24$ ans ; (*iii*) $25-35$ ans ; (*iv*) $36-45$ ans et (*v*) $46-73$ ans.

4.1 Evaluation des recommandations : Precision et Rappel de PERSOREC

Chacun des jeux de données MOVIELENS et BOOKCROSSING a été partitionné en deux échantillons : un échantillon contenant 80% des utilisateurs a été utilisé comme **base d'apprentissage** et un échantillon contenant les 20% d'utilisateurs restants, a été utilisé pour la validation de nos tests (*i.e.,* **base de test**). Pour chaque utilisateur du deuxième échantillon (*i.e.,* utilisateur test), 20% aléatoires de ses tags et ressources sont considérées comme ensemble de test/réponse et 80% comme son ensemble d'apprentissage. Pour chaque utilisateur test, notre algorithme de recommandation génére une liste d'éléments (utilisateurs, tags ou ressources) en se basant sur son ensemble d'apprentissage. Si un élément de la liste de recommandation se trouve également dans l'ensemble de test de cet utilisateur, alors l'élément est considéré comme **pertinent**.

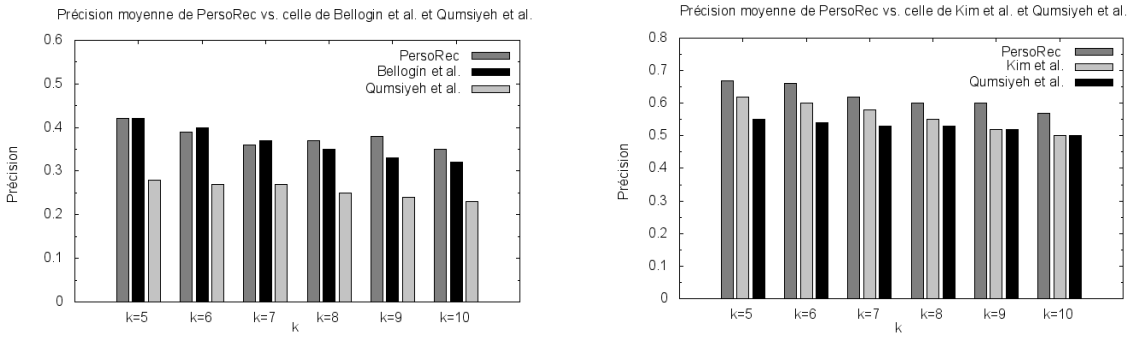


FIGURE 1 – Précision de PERSOREC *vs.* les autres approches pour (**à gauche**) la recommandation de films et (**à droite**) la recommandation de livres.

Évaluer l'efficacité d'un algorithme de recommandation est loin d'être trivial. En premier lieu, parce que différents algorithmes peuvent être meilleurs ou moins bons en fonction du jeu de données sur lequel ils sont appliqués (Herlocker *et al.* (2004)). Néanmoins, pour déterminer l'efficacité d'un système, nous pourrions appliquer les métriques classiques de recherche d'informations : la précision et le rappel (Baeza-Yates & Ribeiro-Neto (1999)). Ces mesures représentent la qualité de la recommandation, c'est-à-dire à quel point les suggestions proposées sont conformes aux intérêts de l'utilisateur. Tout d'abord, la précision détermine la probabilité qu'un élément recommandé soit pertinent. Dans nos expérimentations, nous avons également fait varier le nombre de recommandations proposées dont l'utilisateur peut spécifier le nombre k de réponses les plus pertinentes que le système doit lui retourner. Cela permet surtout d'éviter de submerger l'utilisateur par un grand nombre de réponses en lui retournant que le nombre de réponses les plus pertinentes qu'il souhaite. Dans ce qui suit, nous nous intéressons à la tâche de recommandation de ressources et nous évaluons la précision de notre approche *vs.* les travaux pionniers qui ont un objectif commun avec la nôtre, *i.e.,* ceux de Bellogin *et al.* (Bellogín *et al.* (2013)), Qumsiyeh *et al.* (Qumsiyeh & Ng (2012)) et Kim *et al.* (Kim *et al.* (2011)). Ces approches n'utilisent pas de quatrième dimension mais font appel au profil des utilisateurs comme information complémentaire pour la tâche de recommandation.

Ainsi, la Figure 1 montre les différentes valeurs de précision obtenues par notre algorithme de recommandation *vs.* ses concurrents pour différentes valeurs de k variant entre 5 et 10 sur les deux jeux de données. En général, nos recommandations pour les utilisateurs de MOVIELENS et BOOKCROSSING répondent à leurs attentes. En effet, les recommandations sont pertinentes à 38% et 62%, respectivement, pour MOVIELENS et BOOKCROSSING, ce qui surpasse, pour la plupart des cas, la précision de ses concurrents. Ainsi, pour le jeu de données MOVIELENS, si notre approche fait jeu égal avec celle de Bellogin *et al.* (une précision tout juste meilleure de 2%), l'écart avec l'approche de Qumsiyeh *et al.* est bien plus considérable (une précision plus élevée de 38%). Quant au jeu de données BOOKCROSSING, notre précision est plus grande de, respectivement, 27% et 29% que celles de Kim *et al.* et Qumsiyeh *et al.*. Par ailleurs, les résultats ont montré que PERSOREC atteint ses meilleures performances lorsque la valeur de k est égale à 5. Cela est dû au fait que les cinq premières recommandations correspondent aux besoins des utilisateurs et que lorsque le nombre de recommandations augmente, cela entraîne inévitablement une diminution de la précision étant donné que l'utilisateur choisit moins de

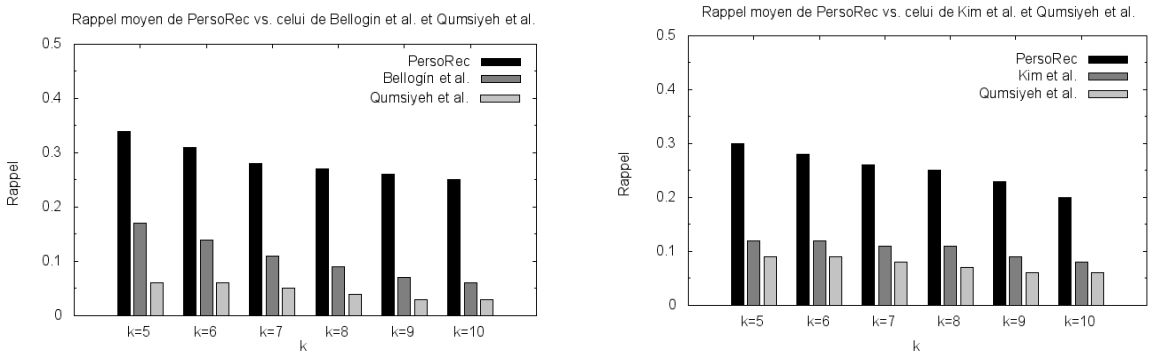


FIGURE 2 – Rappel de PERSOREC *vs.* les autres approches pour (**à gauche**) la recommandation de films et (**à droite**) la recommandation de livres.

ressources que celles qui lui sont recommandées. Quant à la différence entre notre précision et celle des autres approches, nous l'expliquons par le fait que l'utilisation des quadri-concepts améliore les recommandations en suggérant les tags et ressources les plus proches des besoins des utilisateurs. En effet, alors que les travaux connexes se concentrent sur les éléments les plus populaires (livres, films, tags), les quadri-concepts offrent à nos utilisateurs, les tags et les ressources qui ont été partagés en commun par des utilisateurs aux profils proches.

Quant au rappel, il mesure le pourcentage de recommandations pertinentes retournées à l'utilisateur parmi l'ensemble total de recommandations pertinentes. La Figure 2 démontre les différentes valeurs de rappel obtenues par PERSOREC *vs.* les autres approches de la littératures pour différentes valeurs de k allant de 5 à 10 sur les jeux de données MOVIELENS et BOOKCROSSING. Les résultats montrent que notre algorithme surpasse nettement les approches de la littérature. En effet, PERSOREC a une valeur moyenne de rappel égale à 30% sur MOVIELENS vs. 17% pour Bellogin *et al.* et 6% pour Qumsiyeh *et al.*. Quant au jeu de données BOOKCROSSING, notre rappel est, respectivement, $2,50$ et $2,77$ fois meilleur que celui de Kim *et al.* et Qumsiyeh *et al.*. Cette différence de performances démontre que sur l'ensemble total des éléments pertinents, PERSOREC est capable d'en recommander, à ses utilisateurs, une portion plus grande que ses concurrents. Grâce aux quadri-concepts, qui représentent des structures représentatives de la *folksonomie*, PERSOREC recommande donc des éléments partagés par des utilisateurs qui sont susceptibles d'être ensuite partagés par des utilisateurs avec un profil proche.

Le *F1-Score* (ou *mesure F1*) combine à la fois la précision et le rappel qui peut être interprétée comme une moyenne pondérée de ces deux mesures Herlocker *et al.* (2004). La meilleure valeur du *F1-Score* est égale à 1 tandis que la pire valeur correspond à 0. Dans le domaine de la recommandation, cette mesure est un bon indicateur de l'utilité des recommandations. La Figure 3 démontre les différentes valeurs de *F1-Score* obtenues par PERSOREC *vs.* les autres approches de la littératures pour différentes valeurs de k allant de 5 à 10 sur les datasets MOVIELENS et BOOKCROSSING. Les résultats montrent très logiquement que PERSOREC surpasse ses concurrents sur les deux datasets puisque les valeurs de rappel et de précision de PERSOREC sont supérieures à celles des autres approches.

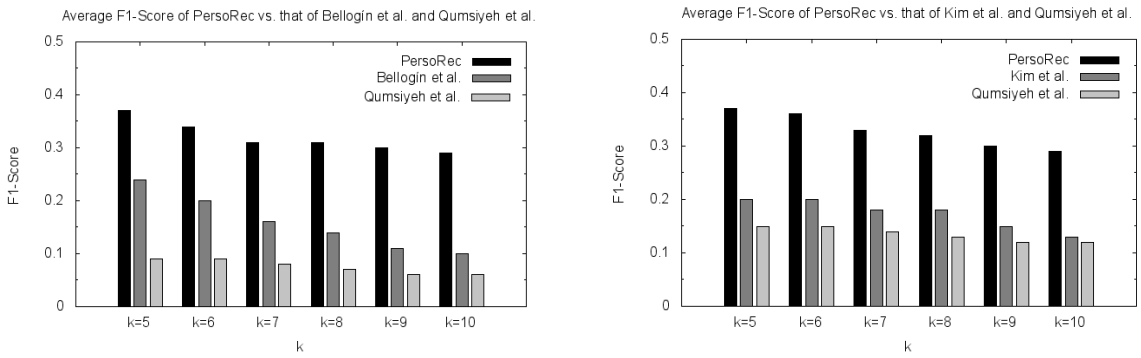


FIGURE 3 – F1-Score de PERSOREC *vs.* les autres approches pour (**à gauche**) la recommandation de films et (**à droite**) la recommandation de livres.

4.2 Evaluation Sociale

Dans ce qui suit, nous étudions l'évaluation sociale de notre système de recommandation. Nous analysons ce qui se passe **après** l'étape de recommandation, *i.e.,* si l'utilisateur cible a vraiment aimé les recommandations et si les utilisateurs (amis) qu'on lui a proposés adoptent le même comportement social. Pour ce faire, nous étudions trois différents cas réels de recommandation avec BOOKCROSSING et un cas réel pour MOVIELENS.

Ainsi, pour le premier jeu de données, nous avons choisi trois utilisateurs avec différents âges et pays : *skinner* (38 ans, New York, USA), *herge* (26 ans, Seixal, Portugal) et *benjamin* (15 ans, Texas, USA). En premier lieu, notre algorithme recommande au premier utilisateur trois différents livres de la franchise *Harry Potter* ainsi que quatre nouveaux amis : *ross* (43 ans, *Illinois*, *USA*), *fran* (54 ans, *California*, *USA*), *emma* (40 ans, *Oregon*, *USA*) et *anna_lucia* (36 ans, *Teheran*, *Iran*). Il s'est avéré plus tard que ces nouveaux amis ont également partagé tous les livres de la franchise *Harry Potter*. De plus, *skinner*, ainsi que ses amis recommandés, ont attribué aux livres recommandés la note de 9 ce qui démontre une réelle appréciation des recommandations. En second lieu, nous avons recommandé à *herge* trois différents livres (*Da Vinci Code*, *Wild Animus* et *The Joy Luck Club*) ainsi que quatre amis agés entre 25 et 35 ans : trois d'entre eux venant des *USA* (*Kansas*, *Wisconsin* et *Virginia*) et le quatrième du *Canada* (*Ottawa*). Cependant, bien qu'il ait partagé les trois livres, *herge* ne les a pas notés, et parmi ses "nouveaux amis", l'un d'eux fut vraiment intéressé par les mêmes livres. En dernier lieu, notre algorithme a généré pour *benjamin* deux livres (*Harry Potter and the Prisonnier of Azkaban* et *Harry Potter and Cup of Fire*) ainsi qu'un nouvel ami, *i.e., baefire* (12 ans, *Illinois*, *USA*). Il se trouve, qu'après la recommandation, les deux utilisateurs ont partagé les livres en leur attribuant la note maximale (10) ce qui indique qu'ils ont vraiment bien apprécié les livres recommandés.

Quant au jeu de données MOVIELENS, notre utilisateur cible est *Bruce* (47 ans, Homme, Educateur). PERSOREC lui a recommandé quatre films : *Star Wars*, *The Return of the Jedi*, *God Father 1 and 2* ainsi que deux nouveaux utilisateurs : *Slioua* (49 ans, Femme, Educatrice) et *Nina* (49 ans, Homme, Educateur). Tout d'abord, nous pouvons remarquer que *Bruce* a apprécié les films recommandés en leur attribuant la note maximale de 5. Ensuite, nous avons noté que ses amis recommandés ont aussi partagé les mêmes films avec une note moyenne de 4 ce qui démontre que Bruce et ses nouveaux amis ont des intérêts communs pour les mêmes films.

Dans ce qui suit, nous proposons plusieurs métriques afin d'évaluer les propriétés de notre système de recommandation. Ces métriques sont définies comme la capacité d'un système de recommandation à suggérer à l'utilisateurs des éléments pertinents mais non populaires.

4.3 Propriétés de PERSOREC

Bien qu'elle soit une tâche cruciale, la recommandation d'utilisateurs, de tags et de ressources s'avère parfois insuffisante pour déployer un bon système de recommandation. Souvent, les utilisateurs peuvent être intéressés par plus qu'une bonne recommandation : découvrir de nouveaux éléments, la diversité des éléments, etc. Ainsi, nous devons identifier l'ensemble de propriétés qui influent sur la réussite d'un système de recommandation (Ricci *et al.* (2011)) :

Couverture de l'Espace Utilisateur. La couverture de l'espace utilisateur est définie comme étant la portion d'utilisateurs pour laquelle le système peut recommander des éléments. Les algorithmes capables de fournir des recommandations à la majorité des utilisateurs sont donc particulièrement appréciés. Ceci dit, PERSOREC est capable de donner des recommandations à tous les utilisateurs de la *folksonomie* indépendamment du fait qu'un utilisateur ait tagué moins d'élements qu'un seuil défini ou le fait qu'il doit avoir un certain nombre d'amis. Dès qu'un utilisateur est ajouté à la *folksonomie*, ses informations personnelles sont suffisantes pour obtenir des recommandations de tags, de ressources et d'utilisateurs (comme démontré dans le critère suivant). Nous avons, par ailleurs, calculé le pourcentage de profils couverts par PERSOREC, ce qui a donné les résultats suivants : 100% de genres (homme et femme), 100% des catégories d'âge, 100% des métiers et 88% des pays [1].

Problème du cold start. Un problème récurrent dans le processus de recommandation est celui du "démarrage à froid" ou cold start (Ricci *et al.* (2011)), *i.e.,* la performance du système vis-à-vis des nouveaux utilisateurs. Dans ce qui suit, nous tentons de répondre à ce problème. Un utilisateur est considéré comme nouveau s'il n'a encore tagué aucune ressource. Contrairement à la majorité des approches de la littérature, PERSOREC ne requiert pas qu'un utilisateur ait partagé un nombre minimum de ressources avant d'être considéré par le système de recommandation. Cela revient au fait que PERSOREC regarde d'abord le profil d'un utilisateur, *i.e.,* ses informations personnelles (âge, profession, etc.) pour lui fournir des recommandations. Ensuite, lorsque ce même utilisateur commence à annoter des ressources avec des tags, le système pourra lui fournir de nouvelles recommandations selon ce qu'il a partagé. Avec cette extension, PERSOREC est maintenant capable de recommander à ses utilisateurs les nouvelles ressources ajoutées et de prendre en compte les nouveaux tags et utilisateurs sans avoir à redémarrer le processus d'extraction des quadri-concepts. Ainsi, notre méthode peut devenir incrémentale.

Diversité. Recommander un ensemble d'éléments qui sont similaires n'est pas aussi utile pour les utilisateurs, qui préfèrent la diversité, *i.e.,* des recommandations qui sont différentes et *distantes*. Par exemple, un utilisateur préférera une recommandation de cinq livres écrits par cinq auteurs différents à une recommandation de cinq livres écrits par un même auteur. Afin de mesurer la diversité de nos recommandations sur le jeu de données BOOKCROSSING, nous utilisons la métrique de distance d (*cf.,* équation 1) qui mesure la distance entre l'élément recommandé et un ensemble d'éléments déjà tagués par l'utilisateur. La métrique de distance d est définie comme suit (Ricci *et al.* (2011)) :

1. À partir de l'ensemble des 13625 pays représentés dans BOOKCROSSING, nous avons évalué la couverture de PERSOREC sur les pays les plus représentées, *i.e.,* les pays présents dans, au moins, 500 quadruplets.

$$d(b, B) = \frac{1 + C_B - C_B.w(b)}{1 + C_B} \quad (1)$$

où *b* est le livre recommandé, B l'ensemble de livres déjà lus par l'utilisateur ciblé, C_B le nombre maximal de livres écrits par un même auteur dans l'ensemble B et $C_{B.w(b)}$ le nombre de livres écrits par l'auteur de b dans l'ensemble B. À noter que la valeur de *d* se tient dans l'intervalle unitaire. Nous utilisons cette métrique de la manière suivante : nous calculons d'abord la distance entre chaque livre recommandé et le reste de la liste des livres recommandés et ensuite, nous calculons la moyenne de ces résultats afin d'obtenir le score de diversité. Nous avons donc obtenu un score de diversité égal à $0,56$ pour l'ensemble des trois utilisateurs (de l'évaluation sociale) avec un score maximal égal à 1 atteint pour le second utilisateur. Ce dernier score s'explique par le fait que nous avons recommandé au deuxième utilisateur trois livres d'auteurs différents, ce qui, en plus de le surprendre, plaît à un utilisateur intéressé par des recommandations diverses.

Scalabilité. L'approche standard pour évaluer la scalabilité d'un système est d'évaluer la complexité de l'algorithme dédié en termes de temps d'exécution ou/et de mémoire requise. Ainsi, nous calculons le temps moyen d'exécution (en millisecondes) de nos recommandations sur les deux jeux de données pour les tâches de recommandation de ressources (dénotée Tâche 1) et de proposition d'utilisateurs (dénotée Tâche 2). Le Tableau 1 affiche tout d'abord le temps d'exécution de PERSOREC sur le jeu de données MOVIELENS qui contient 100000 quadruplets <utilisateur,tag,ressource,profil>. Chaque quadri-concept extrait contient au moins un tag, une ressource et une information de profil, alors que nous faisons varier le support minimum d'utilisateurs (*minsupp_u*), *i.e.,* le nombre minimum d'utilisateurs par quadri-concept. Par exemple, lorsque *minsupp_u* est égal à 6, nous avons 13461 quadri-concepts où chaque concept contient, au moins, 6 utilisateurs. Le Tableau 1 démontre les bonnes performances de PERSOREC pour toutes les valeurs de *minsupp_u*. Tandis que le nombre de quadri-concepts augmente rapidement (de 221 à 13461), le temps d'exécution des recommandations générées par PERSOREC est en moyenne de 2 ms et de 8 secondes pour, respectivement, la première et deuxième tâche. Le nombre total des recommandations (*i.e.,* le nombre d'utilisateurs uniques) est égal à 865 pour la plus petite valeur de *minsupp_u*. Le Tableau 1 affiche également les performances de PERSOREC sur le jeu de données BOOKCROSSING qui contient 762000 quadruplets. Nous avons fait varier le nombre minimum d'utilisateurs par quadri-concept de 30 à 10. Tout d'abord, nous pouvons voir que le nombre maximal de quadri-concepts extraits est égal à 13461 ce qui représente seulement $1,76\%$ de la *folksonomie* ; ce qui démontre, une fois de plus, l'utilité des quadri-concepts, qui sont une représentation réduite de la *folksonomie*. Contrairement à MOVIELENS, le nombre d'utilisateurs uniques, *i.e.,* le nombre total de recommandations, augmente considérablement tandis que le nombre de quadri-concepts fréquents croît légèrement (en raison des valeurs élevées de *minsupp_u*). Cependant, si PERSOREC affiche toujours des bonnes performances pour la tâche de recommandation de ressources où le temps moyen d'exécution est de 28 ms, la tâche de proposition d'utilisateurs devient légèrement plus lente étant donné que chaque quadri-concept contient au moins 10 utilisateurs. Cependant, le temps d'exécution demeure raisonnable à hauteur de 384 secondes en moyenne.

Comparaisons avec les travaux de la littérature. Afin de mettre en lumière les plus-values de notre approche par rapport à ses prédécesseurs, nous surlignons dans le Tableau 2 les différents critères (discutés plus en détails dans la Section 4.3) qu'un système de recommandation

	minsupp_u	$\lvert\mathcal{QC}\rvert$	**# Utilisateurs Uniques**	**Tâche** 1 **(ms)**	**Tâche** 2 **(ms)**
(MOVIELENS)	20	221	526	0,1	2,6
	16	500	605	0,2	3,9
	12	1295	668	0,7	6,1
	8	5123	805	4,0	13,5
	6	13461	865	12,7	23,3
(BOOKCROSSING)	30	553	6789	0,9	149,8
	20	1486	9092	4,9	296,9
	16	2638	10397	13,0	415,5
	12	5698	12239	45,0	542,3
	10	10100	13457	114,7	586,8

TABLE 1 – Temps moyen d'exécution des recommandations de PERSOREC.

doit vérifier Ricci *et al.* (2011). Le point d'interrogation (" ?") dénote qu'une information est manquante dans l'approche et est difficile à vérifier. Nous pouvons remarquer, par exemple, qu'aucune des approches n'offre une recommandations multi-mode (d'utilisateurs, tags et de ressources en même temps). Comme il sera démontré dans la Section 4.3, notre approche diffère de ses prédécesseurs en tenant en compte les nouveaux utilisateurs (les critères *couverture* et *cold start*) en leur fournissant des recommandations sans qu'ils n'aient déjà tagué par le passé. Par ailleurs, si la plupart des approches satisfont le critère de *diversité*, d'autres critères comme la *scalabilité* (le passage à l'échelle) sont difficiles à vérifier étant donné que leurs auteurs n'ont pas donné suffisamment d'informations sur leurs approches.

	Multi-Mode	**Couverture**	**Cold Start**	**Diversité**	**Scalabilité**
Diederich & Iofciu (2006)	Non	Non	Non	Oui	Non
Basile *et al.* (2007)	Non	Non	Non	?	?
Jäschke *et al.* (2007)	Non	Non	Non	Oui	?
Lipczak (2008)	Non	Non	Non	Oui	?
Landia & Anand (2009)	Non	Non	Non	Non	?
De Meo *et al.* (2010)	Non	Non	Non	?	?
Hu *et al.* (2011)	Non	Non	Non	Oui	?
Notre approche	Oui	Oui	Oui	Oui	Oui

TABLE 2 – Les différentes approches de la littérature en bref.

5 Conclusion et Perspectives

Dans ce papier, nous avons introduit une nouvelle approche de recommandation basée sur les concepts quadratiques, et permettant d'offrir un choix personnalisé de tags et de ressources aux utilisateurs. L'évaluation de notre algorithme de recommandation a donné de bons résultats selon les métriques introduites. Nous avons également mené une étude de cas sur six sujets différents afin d'avoir leur feedback et dont les résultats sont disponibles dans la version étendue de notre papier (Jelassi *et al.* (2014)). Parmi les perspectives de nos travaux, nous pouvons considérer d'autres informations comme les reviews, commentaires ou historique de navigation. De plus, pour améliorer les recommandations, il sera utile de disposer d'une évolution temps-réel

des tags et ressources partagés par les utilisateurs, pour cibler encore mieux leurs besoins.

Remerciements. Ce travail est partiellement financé par le projet franco-tunisien PHC Utique 11G141. Nous remercions les relecteurs anonymes pour leurs remarques constructives.

Références

BAEZA-YATES R. & RIBEIRO-NETO B. (1999). *Modern Information Retrieval*. Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc.

BASILE P., GENDARMI D., LANUBILE F. & SEMERARO G. (2007). Recommending smart tags in a social bookmarking system. In *Bridging the Gap between Semantic Web and Web 2.0*, p. 22–29.

BELLOGÍN A., CANTADOR I. & CASTELLS P. (2013). A comparative study of heterogeneous item recommendations in social systems. *Inf. Sci.*, **221**, 142–169.

CERF L., BESSON J., ROBARDET C. & BOULICAUT J.-F. (2009). Closed patterns meet n-ary relations. *ACM TKDD*, **3**, 3 :1–3 :36.

DAS M., THIRUMURUGANATHAN S., AMER-YAHIA S., DAS G. & YU C. (2012). Who tags what ? an analysis framework. *In Proceedings of PVLDB*, **5**(11), 1567–1578.

DE MEO P., QUATTRONE G. & URSINO D. (2010). A query expansion and user profile enrichment approach to improve the performance of recommender systems operating on a folksonomy. *User Modeling and User-Adapted Interaction*, **20**(1), 41–86.

DIEDERICH J. & IOFCIU T. (2006). Finding communities of practice from user profiles based on folksonomies. In *Proceedings of the 1st International Workshop on TEL-CoPs, Crete, Greece*, p. 288–297.

HERLOCKER J. L., KONSTAN J. A., TERVEEN L. G. & RIEDL J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, p. 5–53.

HU J., WANG B. & TAO Z. (2011). Personalized tag recommendation using social contacts. In *Proc. of Workshop SRS'11, in conjunction with CSCW*, p. 33–40.

JÄSCHKE R., HOTHO A., SCHMITZ C., GANTER B. & STUMME G. (2008). Discovering shared conceptualizations in folksonomies. *Web Semantics.*, **6**, 38–53.

JÄSCHKE R., MARINHO L., HOTHO A., LARS S.-T. & STUM G. (2007). Tag recommendations in folksonomies. In *Proc. of the 11th ECML PKDD, Warsaw, Poland*, p. 506–514.

JELASSI M. N., BEN YAHIA S. & MEPHU NGUIFO E. (2013). A personalized recommender system based on users' information in folksonomies. In *Proc. of the 5th Intl. Workshop on Web Intelligence & Communities WI&C at 22nd Intl. WWW conf, Rio de Janeiro, May*, p. 1215–1224.

JELASSI M. N., BEN YAHIA S. & MEPHU NGUIFO E. (2014). Vers des recommandations plus personnalisées dans les folksonomies. *ArXiv e-prints*.

KIM H. K., OH H. Y., GU J. C. & KIM J. K. (2011). Commenders : A recommendation procedure for online book communities. *Electron. Commer. Rec. Appl.*, **10**(5), 501–509.

LAM X. N., VU T., LE T. D. & DUONG A. D. (2008). Addressing cold-start problem in recommendation systems. In *Proc. of the 2Nd ICUIMC*, p. 208–211, New York, NY, USA.

LANDIA N. & ANAND S. (2009). Personalised tag recommendation. *Recommender Systems & the Social Web, New York, NY, USA*, p. 83–86.

LIPCZAK M. (2008). Tag recommendation for folksonomies oriented towards individual users. In *Proc. of the ECML/PKDD Discovery Challenge, Antwerp, Belgium*, p. 84–95.

MIKA P. (2007). Ontologies are us : A unified model of social networks and semantics. *Journal of Web Semantics.*, **5**(1), 5–15.

QUMSIYEH R. & NG Y.-K. (2012). Predicting the ratings of multimedia items for making personalized recommendations. In *SIGIR'12*, p. 475–484.

RICCI F., ROKACH L., SHAPIRA B. & KANTOR P. (2011). *Recommender Systems Handbook*. Springer.

Web sémantique

Swip : une interface Langue Naturelle à SPARQL programmée en SPARQL

Camille Pradel, Ollivier Haemmerlé, Nathalie Hernandez

IRIT, Université de Toulouse le Mirail, Département de Mathématiques-Informatique, 5 allées Antonio Machado, F-31058 Toulouse Cedex 9
{camille.pradel,ollivier.haemmerle,nathalie.hernandez}@univ-tlse2.fr

Résumé : L'approche *Swip* a pour objectif de traduire en SPARQL des requêtes exprimées en langue naturelle en exploitant des patrons de requêtes préalablement définis. Nous présentons ici le module au cœur du système implémentant cette approche qui repose entièrement sur SPARQL. Les traitements mis en œuvre au sein de ce module sont en effet entièrement réalisés sur une base de triplets RDF par l'intermédiaire de requêtes de mise à jour SPARQL. L'implémentation bénéficie ainsi des capacités du moteur SPARQL employé, ce qui permet d'éviter de mettre en place des fonctions de manipulation et d'appariement de graphes, un moteur SPARQL étant justement conçu et optimisé pour ces tâches.

Mots-clés : SPARQL, appariement, application web.

L'approche *Swip* [1] que nous proposons veut fournir aux utilisateurs finals un moyen d'interroger des bases de connaissances sous forme de graphes à l'aide de requêtes exprimées en langue naturelle, et ce dans le but d'éviter à ces utilisateurs de se confronter à la complexité de la formulation d'une requête graphe dans un langage tel que SPARQL.

D'autres travaux visent à générer automatiquement – ou semi-automatiquement – des requêtes formelles à partir de requêtes exprimées sous forme de mots-clés ou de phrases en langue naturelle. L'utilisateur exprime son besoin en information de façon intuitive, sans avoir à connaître le langage de requêtes ou bien le formalisme de représentation de connaissances utilisé dans le système. Certains travaux ont proposé de traduire des requêtes de haut niveau en requêtes formelles dans différents langages comme SeREQL (Lei *et al.*, 2006) ou SPARQL (Zhou *et al.*, 2007; Cabrio *et al.*, 2012). Dans ces systèmes, la génération de requêtes nécessite les étapes suivantes : (i) appariement des mots de la requête aux entités sémantiques de la base de connaissances ; (ii) construction de graphes requêtes liant les entités détectées à l'étape précédente ; (iii) classement des requêtes construites, (iv) sélection de la bonne requête par l'utilisateur. Les approches se sont pour l'instant focalisées sur des problèmes précis : optimiser l'étape d'appariement en exploitant des ressources externes comme Wordnet ou Wikipedia (Lei *et al.*, 2006; Wang *et al.*, 2008; Cabrio *et al.*, 2012), optimiser l'indexation et l'exploration de la base de connaissances pour la construction de la requête graphe (Zhou *et al.*, 2007), améliorer le classement des requêtes candidates (Wang *et al.*, 2008), optimiser l'identification de relations à l'aide de patrons textuels (Cabrio *et al.*, 2012), ou encore mettre en œuvre un dialogue avec l'utilisateur pour raffiner l'interprétation de la requête utilisateur (Lehmann & Bühmann, 2011; Unger *et al.*, 2012).

L'originalité de notre approche réside aussi bien dans la démarche mise en place pour interpréter la requête que dans son implémentation faisant une utilisation poussée de SPARQL.

1. http://swip.univ-tlse2.fr/SwipWebClient/welcome.html

La section 1 présente les originalités de l'approche et de son déploiment. La section 2 détaille les premières étapes du processus afin de donner une idée précise de son implémentation. La section 3 discute des avantages et inconvénients de cette approche.

1 Cadre

Nous définissons ici le cadre requis pour la mise en œuvre de notre approche. Nous présentons en 1.1 les originalités de l'approche, en 1.2 les ontologies que nous avons conçues afin de fournir un cadre logique à l'implémentation et en 1.3 les quelques moyens logistiques nécessaires.

1.1 Originalités

Une des originalités de *Swip* se caractérise par l'utilisation de *patrons de requêtes* préétablis pour guider le processus d'interprétation. Nos travaux se fondent en effet sur le postulat selon lequel, dans les applications réelles, les requêtes formulées par les utilisateurs sont pour l'essentiel des variations autour de quelques familles typiques de requêtes. Chaque patron de requêtes représente une de ces familles de requêtes. Les patrons de requêtes sont constitués d'un graphe représentant le besoin en informations couvert par le patron et faisant référence à des ressources de la base de connaissances cible, et d'un modèle de phrase descriptive permettant de générer des phrases en langue naturelle présentées à l'utilisateur.

Le processus d'interprétation de la requête utilisateur est décomposé en deux étapes principales, avec un résultat intermédiaire qui est la *requête pivot*. La requête pivot consiste en une première interprétation de la requête utilisateur dans laquelle le besoin en information est exprimé sous une forme proche d'une requête par mots-clés, mais dans laquelle il est possible d'exprimer des relations entre les mots-clés (Pradel *et al.*, 2011). Par exemple, la requête pivot `?"person": "produce"= "In Utero". "In Utero": "album"` est obtenue lors de l'interprétation de la requête en langue naturelle "Who produced the album In Utero ?" (tiré du jeu d'entraînement de la compétition QALD-3 [2]). Cette organisation présente deux avantages principaux : elle permet de représenter les informations importantes issues de l'analyse syntaxique de la requête utilisateur, et elle facilite la mise en œuvre du multilinguisme dans notre approche. En effet, la requête pivot est un format intermédiaire indépendant de la langue, et peut donc être traitée de la même façon quel que soit le langage employé par l'utilisateur. Ainsi, pour adapter notre approche à un nouveau langage, il suffit de modifier la première étape.

Cette première grande étape consiste en l'interprétation de la requête utilisateur et sa traduction en requête pivot. Pour cela, des opérations classiques de traitement automatique des langues (identification des entités nommées, détermination des catégories grammaticales, analyse de dépendances) sont appliquées à la requête en langue naturelle, puis des règles de transformation préétablies sont utilisées pour générer la requête pivot à partir de l'arbre de dépendances de la phrase. Ce premier module est architecturé de façon classique pour une application web : des services web effectuent les sous-tâches basiques du processus d'interprétation et sont exploités par un *workflow*, lui-même accessible sous forme de service web. Chaque service exploite en

2. http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/index.php?x=task1&q=3

entrée le résultat du ou des services situés en amont dans le *workflow*. Ce premier module et l'ensemble des services web qui le composent ont été présentés dans (Pradel *et al.*, 2013b,a).

Dans la seconde étape, les patrons de requêtes sont associés à la requête pivot pour obtenir une liste d'interprétations possibles de cette requête (et donc de la requête utilisateur d'origine) qui sont ensuite ordonnées en fonction de leurs pertinences supposées. Les interprétations peuvent ainsi être soumises à l'utilisateur sous forme de phrases descriptives générées à partir des modèles de phrases descriptives de chaque patron. L'implémentation de ce module est originale et présente selon nous une réelle nouveauté. En effet, dans cette implémentation, les traitements permettant les appariements et associations décrits dans (Pradel *et al.*, 2011) sont entièrement réalisés sur une base de triplets RDF par l'intermédiaire de requêtes de mise à jour SPARQL. Cette approche permet d'éviter l'implémentation de fonctions de manipulation et d'appariement de graphes et d'exploiter les capacités d'un moteur SPARQL qui est justement conçu et optimisé pour apparier des graphes. Ainsi, cette tâche est entièrement réalisée via des requêtes SPARQL. Le résultat de chaque étape est systématiquement enregistré dans la base de connaissances, via les fonctionnalités de mise à jour de SPARQL (*SPARQL updates*), ce qui les rend directement exploitables dans l'étape suivante. A notre connaissance, aucune approche ne repose sur une telle exploitation de SPARQL. Ce second module est présenté dans le reste de l'article.

1.2 Ontologies de patrons et de requêtes

Les triplets manipulés et générés lors du processus d'interprétation exploitent le vocabulaire de deux ontologies construites par nos soins et accessibles depuis la page d'accueil du système *Swip* [3].

L'ontologie *patterns*, présentée dans (Pradel *et al.*, 2012), permet de modéliser des patrons de requête et facilite ainsi la gestion, le partage et l'évolution de ces patrons. Les données RDF représentant les patrons de requêtes au format spécifié par cette ontologie trouvent ici une nouvelle utilité étant donné que, une fois insérées dans la base de triplets, elles sont exploitées telles quelles pour interpréter la requête utilisateur.

L'ontologie *queries* permet quant à elle de modéliser des requêtes pivot et définit les structures communes pour représenter les résultats intermédiaires et finals du processus d'interprétation. Nous ne la détaillons pas ici, étant donné qu'elle a été conçue selon les mêmes principes que l'ontologie *patterns* et se révèle plus simple que cette dernière. Comme nous l'expliquons dans (Pradel *et al.*, 2013a), une requête pivot est un ensemble de sous-requêtes, chaque sous-requête étant un 1/2/3-uplet (ensemble de un, deux ou trois éléments) de requête. Les classes et propriétés de l'ontologie *queries* reflètent logiquement cette structure ; leurs noms sont explicites, et un exemple de requête pivot et une partie des résultats de son interprétation exprimés en RDF et fondés sur cette ontologie est montré plus bas.

Cette ontologie intègre également des axiomes permettant l'inférence de nouveaux triplets. Cependant, ces possibilités d'inférence ne sont ici pas exploitées pour des raisons de performances : contrairement à la description des patrons qui est statique (mis à part dans les cas exceptionnels et ponctuels d'évolution ou d'ajout de patrons), les triplets générés au cours de l'interprétation évoluent en continu, et il n'est donc pas possible d'inférer les connaissances

3. http://swip.univ-tlse2.fr/SwipWebClient/welcome.html

liées à l'ontologie *queries* en pré-traitement. Or, il est très coûteux d'activer un raisonneur sur un jeu de données contenu dans un serveur SPARQL, particulièrement lorsque ce jeu de données fait l'objet de nombreuses mises à jour. C'est pourquoi dans notre implémentation, l'ensemble des triplets exploités dans la suite du processus sont générés explicitement à chaque étape et ce, même s'ils sont redondants et pourraient être obtenus à partir d'autres triplets et des connaissances ontologiques.

1.3 Infrastructure

Le processus d'interprétation est intégralement fondé sur des requêtes SPARQL. La principale infrastructure nécessaire à la mise en œuvre de cette approche est donc un moteur SPARQL qui recevra les requêtes de mises à jour émises par l'interface employée par l'utilisateur. Les triplets générés par chaque requête, exploitant le vocabulaire de l'ontologie *Queries*, seront ajoutés à la base de triplets de ce moteur. Les expérimentations que nous avons menées exploitent *Fuseki* [4], un serveur SPARQL intégrant le moteur de requêtes *ARQ* [5].

Pour mener à bien le processus d'interprétation, il est évidemment nécessaire d'avoir accès au jeu de donnés ciblé par la requête utilisateur et aux patrons de requêtes (préalablement traduits en RDF selon le vocabulaire défini dans l'ontologie *Patterns*). Ces éléments peuvent être directement inclus dans la base de triplets du serveur SPARQL ou simplement, grâce aux fonctionnalités de fédération de SPARQL 1.1, distribués sur le web de données à partir du moment où ils sont accessibles via SPARQL. La figure 1 illustre un scénario possible de déploiement.

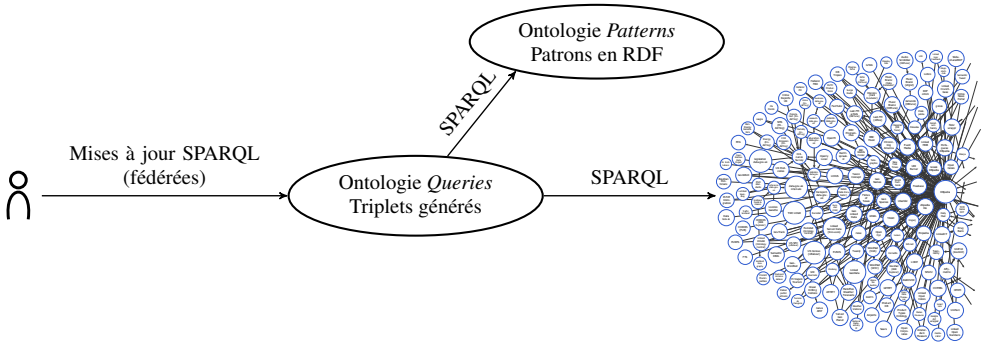


FIGURE 1 – Exemple de déploiement du système *Swip*.

2 Description des premières étapes

Cette section détaille les premières étapes du processus d'interprétation de la requête pivot. Par souci de concision, les étapes suivantes sont présentées plus grossièrement dans la sous-section 2.4. La figure 2 montre toutes les étapes de cette implémentation, chaque étape correspondant à une requête `UPDATE` ou `ASK` (qui permet d'émuler une boucle, comme expliqué plus bas en 2.4) ; ces requêtes sont données sur la page web de présentation de l'approche *Swip* [6].

4. http://jena.apache.org/documentation/serving_data/
5. http://jena.apache.org/documentation/query/
6. http://swip.univ-tlse2.fr/SwipWebClient/welcome.html

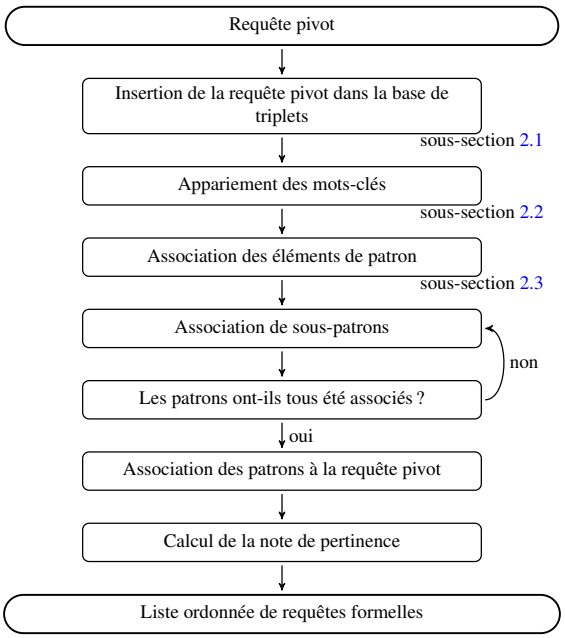


FIGURE 2 – Détail des étapes de formalisation de la requête pivot.

2.1 Insertion de la requête pivot dans la base de connaissances

Dans un premier temps, le système génère un URI qui est unique pour chaque ensemble de requêtes pivot équivalentes. Si cet URI existe déjà dans la base de connaissances, la requête précédente a déjà été traitée et les résultats sauvegardés peuvent être exploités tels quels. Sinon, le système génère un graphe RDF (fondé sur l'ontologie *queries* introduite plus haut) représentant la requête pivot et l'insère au moyen d'une requête de mise à jour SPARQL (`INSERT DATA`) dans la base de connaissances avant de commencer l'interprétation de la requête. Le graphe RDF produit pour la requête `?"person": "produce"= "In Utero". "In Utero": "album"` (qui est la requête pivot issue de la requête "Who produced the album In Utero ?") est montré dans la partie gauche de la figure 3. Cette figure montre les données RDF initiales ainsi que les résultats des mises à jour SPARQL qui sont effectuées au cours de l'étape d'appariement (cf. sous-section 2.2). Les ressources RDF sont représentées dans des nœuds ovales, les littéraux dans des rectangles, et les propriétés sont logiquement matérialisées par des arcs étiquetés. Pour des besoins de lisibilité, les types des littéraux ne sont pas montrés et les classes auxquelles appartiennent les ressources ne sont montrées que lorsque cela aide à la compréhension ; l'URI de la classe d'un nœud est alors indiqué en-dessous de ce nœud.

2.2 Appariement des éléments de la requête à la base de connaissances

La première étape de l'interprétation de la requête pivot consiste à apparier chaque élément de la requête pivot aux entités de la base de connaissances (classes, propriétés et instances) ou aux types de littéraux, en associant à chaque appariement une note de confiance qui représente la qualité supposée de l'appariement et sa vraisemblance.

L'*appariement des mots-clés* est effectué en calculant une mesure de similarité entre les

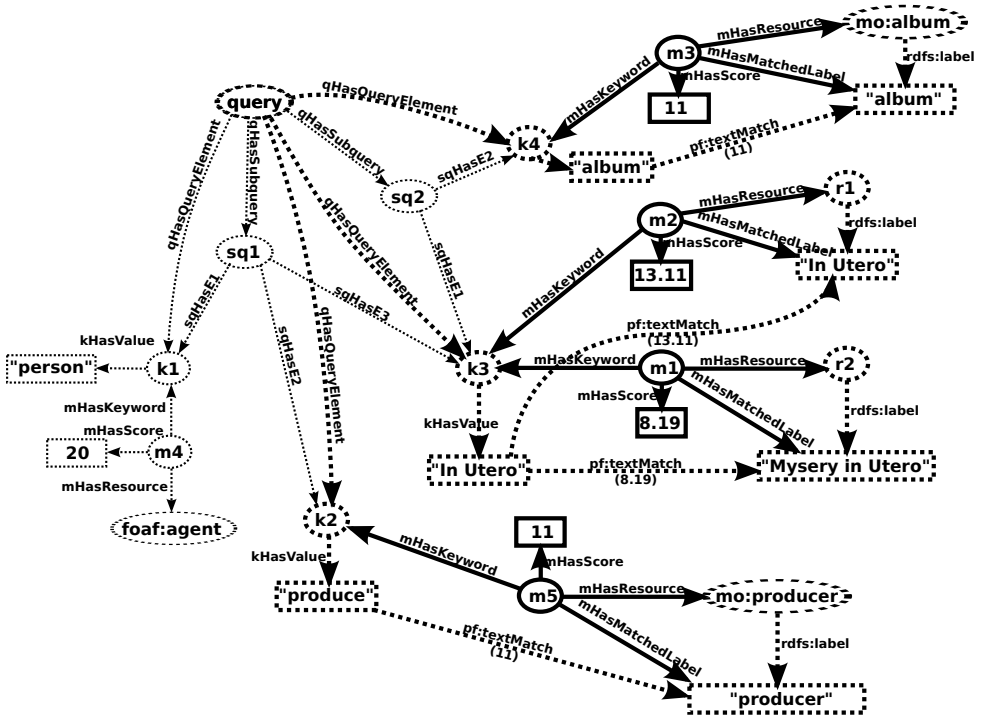


FIGURE 3 – L'étape d'appariement sur notre requête pivot exemple.

mots-clés de la requête pivot et les étiquettes des ressources de la base de connaissances. Pour cela, nous utilisons *LARQ* [7], une extension de SPARQL proposée par le moteur SPARQL *ARQ* qui permet d'exploiter les capacités d'indexation et de recherche dans du texte du moteur de recherche *Apache Lucene* [8]. Cette extension introduit un nouvel élément de syntaxe qui permet de déterminer les littéraux ressemblant à une chaîne de caractères donnée et une valeur représentant le niveau de ressemblance, appelée *score Lucene* : le "triplet" `(?lit ?score) pf:textMatch '+text'` lie à la variable `?lit` tous les littéraux qui sont similaires à la chaîne de caractères `text` et à la variable `?score` le *score Lucene* correspondant. La requête SPARQL intégrale utilisée pour mettre en œuvre cette étape d'appariement est montrée dans la figure 4.

La figure 3 montre un sous-ensemble des appariements obtenus par l'exécution de cette requête et les instances d'appariements (classe `queries:Matching`) générées. Le graphe avant l'exécution de la requête est tracé en pointillés ; la partie de ce graphe qui est appariée au motif de graphe de la clause `WHERE` est en gras, et les ressources et triplets insérés dans la base de connaissances sont en traits pleins. Les figures suivantes utilisent les mêmes conventions graphiques.

Il est important de noter que, bien que cette étape telle que nous l'avons décrite soit mise en œuvre à l'aide d'une extension (non standard) de SPARQL qui la rend peu portable, une alternative peut facilement être implémentée en utilisant des fonctions standard de SPARQL telles que les fonctions de chaînes de caractères `REGEX` et `CONTAINS`, ou encore une simple comparai-

7. http://jena.apache.org/documentation/larq/
8. http://lucene.apache.org/

```
INSERT
{
  ?matchUri a queries:Matching;
            queries:matchingHasKeyword ?keyword;
            queries:matchingHasResource ?r;
            queries:matchingHasScore ?score;
            queries:matchingHasMatchedLabel ?l.
  ?keyword queries:keywordAlreadyMatched "true"^^xsd:boolean.
}
WHERE
{
  <[queryUri]> queries:queryHasQueryElement ?keyword.
  ?keyword a queries:KeywordQueryElement;
           queries:queryElementHasValue ?keywordValue.
  FILTER NOT EXISTS { ?keyword queries:keywordAlreadyMatched "true"^^xsd:boolean. }
  (?l ?score) pf:textMatch (?keywordValue 6.0 5).
  ?r rdfs:label ?l.
  BIND (UUID() AS ?matchUri)
}
```

FIGURE 4 – Mise à jour SPARQL utilisée pour apparier les mots-clés de la requête pivot.

son de chaînes de caractères. Cette version standard présenterait néanmoins des performances dégradées, l'appariement entre chaînes de caractères étant établi de façon exacte.

2.3 Association des éléments de patron aux éléments de requête

Avant de pouvoir associer l'intégralité des patrons à la requête pivot, une première tâche consiste à déterminer pour chaque élément de patron toutes les associations possibles aux éléments de la requête utilisateur et leur note de confiance respective. Ces associations sont appelées *associations d'élément*.

Cette étape consiste à créer une association entre un élément de requête et un élément de patron quand cet élément de requête a été apparié à une ressource qui est liée d'une façon ou d'une autre à l'élément de patron ciblé. Nous définissons plusieurs cas permettant d'établir un lien entre une ressource appariée r et un élément de patron ciblé t :

1. r est une sous-classe de t (ce cas comprend celui où r est la classe t elle-même),
2. r est une sous-propriété de t (ce cas comprend celui où r est la propriété t elle-même),
3. r est une instance de t (ce cas comprend celui des associations d'élément instance-classe, présenté plus bas),
4. r fait référence au même type de littéral que t.

À chaque association d'élément est assignée une note de confiance qui est la même que celle de l'appariement impliqué. La figure 5 montre l'instanciation de quelques appariements via une requête de mise à jour SPARQL. L'élément de patron `cd_info_element5` qui cible la classe `mo:Record` est associé deux fois au mot-clé `k1` (`"In Utero"`) qui a été préalablement apparié à deux instances de cette même classe (cas 3). L'élément de patron `cd_info_element4` qui cible la propriété `mo:producer` est associé au mot-clé `k2` (`"produce"`) qui a été préalablement apparié à cette même propriété (cas 2).

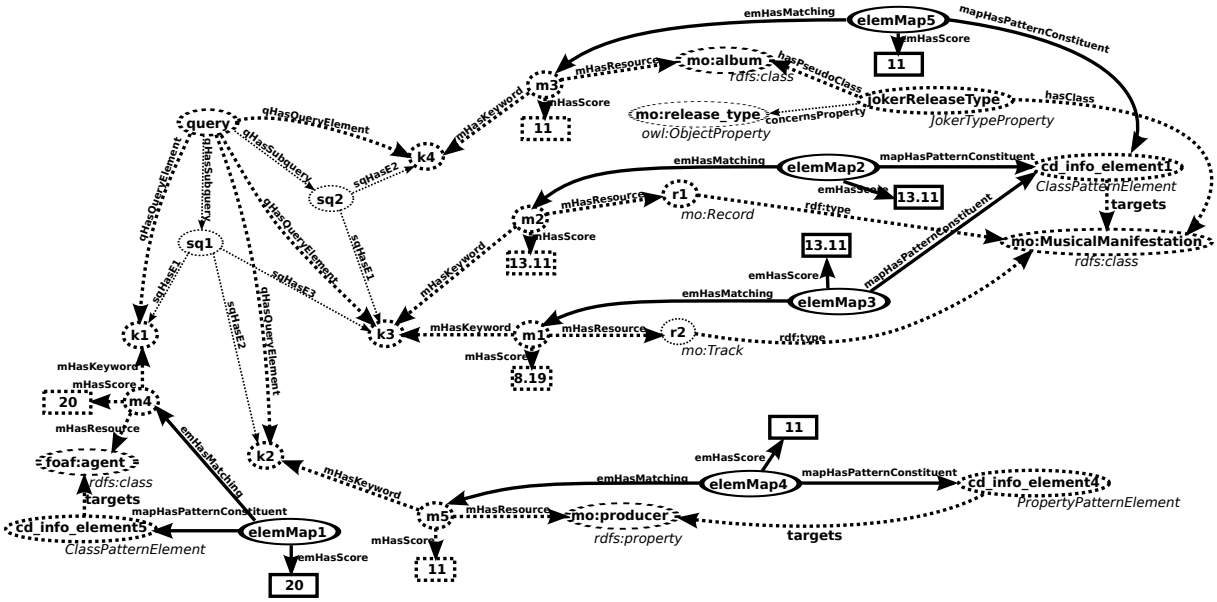


FIGURE 5 – L'étape d'association d'élément sur notre requête pivot exemple.

2.3.1 Le cas de propriétés "*Type"

L'instanciation de l'association d'élément `elemMap5` est issue d'une extension du premier cas établi plus haut. Cette extension est due à l'observation d'un choix de modélisation récurrent fait par certains développeurs d'ontologies, qui consiste à classifier les instances d'une classe c en définissant une propriété d'objet avec pour domaine c, une classe (énumérée) c' qui représente le co-domaine de cette propriété, et des instances de c' qui représentent les différentes façons de classifier les instances de c. Par exemple, dans l'ontologie *music* (Raimond *et al.*, 2007), les instances de la classe `mo:Record` peuvent être impliquées en tant que sujet dans un triplet avec pour prédicat `mo:releaseType` et pour objet une instance de la classe (`mo:Album`, `mo:Single`, `mo:Live`, `mo:Soundtrack`...). Il nous semble que ce choix, bien qu'il ait probablement été guidé par certaines exigences, n'est pas pertinent car il ignore le mécanisme de classification proposé par RDFS et OWL (à savoir le typage d'instances à l'aide de classes) pour exprimer des connaissances qui sont en fait bel et bien une classification. Nous appuyons notre critique en avançant deux éléments que nous considérons comme symptomatiques de ce travers et que l'on retrouve dans notre exemple :

– deux termes qui ont le même statut dans une phrase en langue naturelle (comme par exemple les termes "songs" et "soundtracks" qui sont utilisés de la même façon dans les requêtes "Give me all songs by Aretha Franklin" et "Give me all soundtracks composed by John Williams," toutes deux extraites du jeu de données de la compétition QALD-3) sont modélisés de manière différente dans l'ontologie, le premier par une classe, le second par une instance de la classe `mo:ReleaseType` ;
– la façon dont les developpeurs de l'ontologie eux-mêmes ont nommé la propriété d'objet mise en cause trahit la nature réelle de cette propriété ; en effet, une propriété dont le nom finit par `Type` sera très probablement utilisée pour typer des instances et devrait donc en tant que telle être une sous-propriété de `rdf:type`, or ce n'est pas le cas ; cette seconde

observation est à l'origine du nom que nous avons donné à ce type de propriétés : les propriétés "*Type" (*wildcard type properties* en anglais, d'où l'URI de la classe utilisée plus bas).

Notre approche permet de pallier ce manque de généricité en identifiant explicitement ces cas. Dans la figure 5, l'instance `wildcardReleaseType` de type `WildcardTypeProperty` exprime le fait que la ressource `mo:album` doit être considérée durant le processus d'association comme une sous-classe de `mo:Record`, ce qui permet l'association de l'élément qualifiant `cd_info_element1` au mot-clé `k4` ; elle spécifie aussi que la propriété de typage correspondante (qui devra être utilisée dans la requête SPARQL finale) est dans ce cas `mo:release_type` au lieu du classique `rdf:type`.

2.3.2 Les associations d'élément instance-classe

Un dernier type d'association d'élément peut être produit sur la base des précédents. Ces associations, appelées *associations d'élément instance-classe* sont issues de l'observation de la manière dont les utilisateurs, quand ils s'expriment en langue naturelle, spécifient souvent un terme faisant référence à une instance par un autre terme faisant référence à une classe à laquelle appartient cette instance. On trouve de nombreux exemples dans les requêtes en langue naturelle de la compétition QALD-3 : "the band Dover", "the album In Utero", "the song Hardcore Kids"…

D'après (Pradel *et al.*, 2013a), ce type de formulation se traduit en langage pivot par une sous-requête binaire, composée du mot-clé faisant référence à l'instance qualifié par le mot-clé faisant référence à la classe ; par exemple, "the album In Utero" devient `"In Utero":"album"`. Nous utilisons un type particulier d'association d'élément pour prendre en compte ce cas ; une association de ce type, appelée *association d'élément instance-classe*, associe un élément de patron à deux mots-clés. Sa valeur de confiance est égale à la somme des valeurs de confiance des deux appariements pris en compte.

L'implémentation que nous proposons gère de façon simple ce cas particulier. La figure 6 poursuit le même exemple et illustre l'instanciation d'une association d'élément instance-classe. `elemMap1` et `elemMap4` instanciées préalablement associent `cd_info_element1` respectivement à `k3` (`"In Utero"`) et `k4` (`"album"`), et la ressource appariée par `k3` est une instance (lorsque l'on considère la remarque précédente sur les propriétés "*Type") de la ressource appariée par `k4` ; cela permet d'instancier une nouvelle association d'élément `elemMap6` associant l'élément de patron considéré aux deux éléments de requête. Son score est la somme des notes de confiance des deux associations d'élément originellement impliquées.

2.4 Étapes suivantes

Nous ne détaillons pas autant les étapes suivantes du processus d'interprétation par souci de brièveté. Encore une fois, les requêtes SPARQL utilisées et l'ontologie organisant les triplets générée sont disponibles sur la page d'accueil du système *Swip*.

Comme nous l'avons présenté dans (Pradel *et al.*, 2011), l'étape d'appariement des sous-patrons implique des opérations complexes, comme des combinaisons d'éléments d'un ensemble ou des produits cartésiens entre des ensembles dont le nombre ne peut être déterminé à l'avance. De plus, l'association d'un patron nécessite de commencer par associer les sous-

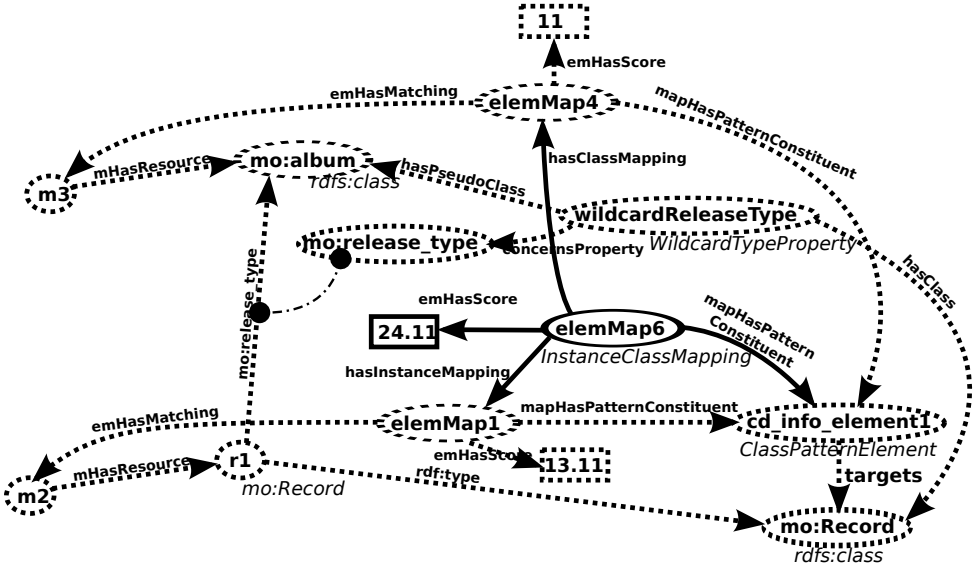


FIGURE 6 – Une association d'élément instance-classe sur notre requête pivot exemple.

patrons les plus simples qui ne contiennent aucun autre sous-patron, puis les sous-patrons qui les contiennent directement, et ainsi de suite jusqu'à avoir associé le patron lui-même. Ces besoins sont traditionnellement résolus en programmation impérative par une structure de contrôle de type *boucle* (*tant que*), elle-même permise par une fonctionnalité de saut conditionnel (*si*).

Une simple succession de mises à jour SPARQL ne peut répondre à ces besoins. Nous avons donc dû ajouter la possibilité d'effectuer un saut conditionnel dans notre implémentation. Ce saut conditionnel est mis en œuvre de manière très simple au travers d'une requête SPARQL `ASK` : deux embranchements sont possibles à l'issue de l'exécution d'une telle requête et celui réellement emprunté dépend du résultat de cette exécution. Comme on peut le voir sur la figure 2, cette méthode est utilisée à la fin de l'étape d'association des sous-patrons, au niveau de la requête (`ASK`) "*Les patrons ont-ils tous été associés ?*" Si la réponse est non (`FALSE`), l'embranchement suivi renvoit à l'exécution de requêtes en amont dans le processus pour revenir au test plus tard. Si la réponse est oui (`TRUE`), alors on passe aux étapes suivantes. Ainsi, une boucle est créée permettant d'assurer que tous les sous-patrons ont été associés avant de passer à la suite.

3 Critique de l'implémentation

L'architecture décrite dans cet article présente pour nous de nombreux avantages. L'un des plus évidents est la facilité d'utilisation d'un système de cache. En effet, étant donné que le résultat de chaque étape de traitement est inséré dans l'entrepôt RDF, il est alors très simple de réexploiter les données générées précédemment. Par exemple, comme expliqué dans la sous-section 2.1, le système Swip se rend compte qu'une requête entrante a déjà été traitée lorsque son URI est déjà présent dans la base de triplets et le résultat de son interprétation peut être directement retourné. De même, l'appariement d'un mot-clé donné n'est réalisé qu'une seule fois pour l'ensemble des requêtes contenant ce même mot-clé.

De plus, cette architecture permet un asynchronisme total et naturel entre le client et le serveur : une fois que l'URI de la requête est construit, il est retourné au client qui peut alors mettre à jour les résultats intermédiaires de l'interprétation en cours progressivement et indépendamment du serveur. Pour cela, il requête directement à l'aide de SPARQL l'entrepôt RDF où sont enregistrés ces résultats.

Cette approche est également homogène et cohérente dans la mesure où les données d'entrée, les données intermédiaires et les données de sortie sont enregistrées et manipulées via des standards. Même la configuration du système et les ajustements sont réalisés de cette manière ; par exemple, les cas identifiés de propriétés "*Type" (présentés plus haut) et certains appariements utiles (c'est-à-dire des appariements de mots-clés au niveau de l'ontologie qui ne pourraient pas être obtenus par mesure de similarité entre chaînes de caractères, comme par exemple entre le mot-clé `"husband"` et la propriété `rel:spouseOf`) sont directement exprimés en RDF, insérés dans la base de triplets et exploités tels quels au cours du processus d'interprétation.

Nous attirons également l'attention du lecteur sur le fait que, bien que la première grande étape du processus d'interprétation (traduction de la langue naturelle vers le langage pivot) ait été implémentée d'une façon plus "traditionnelle," en utilisant des services web, cela pourrait maintenant être fait différemment en exploitant la récente initiative *NLP2RDF* (Hellmann *et al.*, 2013) qui veut fournir un format commun pour les données de sortie des outils de traitement automatique du langage les plus populaires, ce format appelé *NIF (NLP Interchange Format)* étant totalement intégré au cadre du web sémantique.

Enfin, cette architecture permet de déployer le système *Swip* de façon distribuée, ce qui facilite le passage à grande échelle. En effet, pour fonctionner, *Swip* a simplement besoin d'avoir accès via SPARQL aux données à interroger et aux patrons de requêtes. Grâce aux récentes fonctionnalités de fédération de SPARQL (*federated SPARQL* (Prud'hommeaux & Buil-Aranda, 2013)), les données peuvent être regroupées sur un seul serveur SPARQL ou réparties sur plusieurs, et les patrons peuvent eux aussi être regroupés ou répartis, rassemblés avec la base de connaissances qu'ils concernent ou isolés. Cette architecture exploitant des standards est donc très souple et permet de nombreuses variations. De plus, ce sont les serveurs SPARQL qui effectuent les traitements, ce qui rend le programme initial (celui qui émet le requêtes SPARQL) très léger. On peut ainsi imaginer que ce programme, exécuté côté client, met en branle, via la fédération de SPARQL, de nombreux serveurs répartis sur le web et les orchestre pour traiter l'interprétation d'une requête.

Nous retenons également deux inconvénients majeurs qui viennent ternir le tableau. Le premier est le manque de contrôle sur l'exécution des requêtes SPARQL et par conséquent sur les performances générales du système. Le serveur SPARQL est utilisé comme une boîte noire et son efficacité influence directement celle du processus d'interprétation. L'expérience a montré que le serveur ARQ n'est pas performant pour traiter de grosses requêtes (plus de vingt triplets répartis dans plusieurs sous-requêtes) et que la division de ces requêtes en une série équivalente de requêtes successives permet une nette amélioration des performances.

De plus, SPARQL est encore une recommandation relativement jeune qui, malgré les nouvelles fonctionnalités apportées par SPARQL 1.1, propose un panel de fonctions assez limité. Par exemple, il y a très peu de fonctions arithmétiques ; seules les plus basiques sont supportées, ce qui n'est pas le cas de la fonction puissance. En conséquence, une solution dégradée a dû être trouvée pour le calcul de la note de pertinence finale qui impliquait initialement cette fonction.

4 Perspectives

Nous voulons explorer les nouvelles pistes ouvertes par l'implémentation décrite dans ce papier. Nous pensons en effet que l'approche que nous avons utilisée peut être généralisée et utilisée pour d'autres applications : il serait ainsi possible d'implémenter tous types d'algorithmes, et en particulier des algorithmes manipulant des graphes. De plus, pour les raisons exposées en 3, cette approche semble parfaitement adaptée au développement d'applications web et pourrait très bien s'adapter aux cadres proposés visant à intégrer les API web au web de données, comme *RDF-REST* (Champin, 2013).

Nous voudrions donc formaliser cette nouvelle forme de programmation exploitant les mises à jour SPARQL et la comparer à des paradigmes proches, comme la programmation dirigée par les données (*data-driven programming*), ou des approches plus pratiques exploitant des bases de données, comme PL/SQL. Ces travaux nous mèneraient certainement à proposer une extension à SPARQL permettant le saut conditionnel dans le but de faire de ce langage un langage de programmation particulièrement adapté à l'implémentation d'algorithmes impliquant des graphes.

Références

CABRIO E., COJAN J., APROSIO A. P., MAGNINI B., LAVELLI A. & GANDON F. (2012). QAKiS : an open domain QA system based on relational patterns. In *International Semantic Web Conference (Posters & Demos)*, volume 914.

CHAMPIN P.-A. (2013). RDF-REST : a unifying framework for web APIs and linked data.

HELLMANN S., LEHMANN J., AUER S. & BRÜMMER M. (2013). Integrating NLP using linked data.

LEHMANN J. & BÜHMANN L. (2011). AutoSPARQL : let users query your knowledge base. In *The Semantic Web : Research and Applications*, p. 63–79. Springer.

LEI Y., UREN V. & MOTTA E. (2006). Semsearch : A search engine for the semantic web. In *Managing Knowledge in a World of Networks*, p. 238–245. Springer.

PRADEL C., HAEMMERLÉ O. & HERNANDEZ N. (2011). A semantic web interface using patterns : the SWIP system. In *Proceedings of GKR 2011*, p. 172–187, Barcelona, Spain : Croitoru et al.

PRADEL C., HAEMMERLÉ O. & HERNANDEZ N. (2012). Des patrons modulaires de requêtes SPARQL dans le système SWIP. In *23es Journées Francophones d'Ingénierie des Connaissances*, p. 621–636.

PRADEL C., HAEMMERLÉ O. & HERNANDEZ N. (2013a). Natural language query interpretation into SPARQL using patterns. In *COLD@ISWC2013*, Sydney (Australia).

PRADEL C., HAEMMERLÉ O. & HERNANDEZ N. (2013b). SWIP at QALD-3 : results, criticisms and lesson learned. Valencia, Spain.

PRUD'HOMMEAUX E. & BUIL-ARANDA C. (2013). SPARQL 1.1 federated query.

RAIMOND Y., ABDALLAH S. A., SANDLER M. B. & GIASSON F. (2007). The music ontology. In *ISMIR*, p. 417–422.

UNGER C., BÜHMANN L., LEHMANN J., NGONGA NGOMO A.-C., GERBER D. & CIMIANO P. (2012). Template-based question answering over RDF data. In *Proceedings of the 21st international conference on World Wide Web*, p. 639–648.

WANG H., ZHANG K., LIU Q., TRAN T. & YU Y. (2008). Q2semantic : A lightweight keyword interface to semantic search. In *The Semantic Web : Research and Applications*, p. 584–598. Springer.

ZHOU Q., WANG C., XIONG M., WANG H. & YU Y. (2007). SPARK : adapting keyword query to semantic search. In *The Semantic Web*, p. 694–707. Springer.

SPARQL Template : un langage de Pretty Printing pour RDF

Olivier Corby[1] & Catherine Faron-Zucker[2]

[1] INRIA Sophia-Antipolis Méditerranée, 06900 Sophia Antipolis, France
olivier.corby@inria.fr

[2] Univ. Nice Sophia Antipolis, CNRS, I3S, UMR 7271, 06900 Sophia Antipolis, France
faron@i3s.unice.fr

Abstract : RDF est un langage de représentation de connaissances basé sur des graphes étiquetés, conçu par le W3C pour le Web sémantique et le Web des données. En tant que langage d'échange pivot, il peut être utilisé pour représenter des arbres de syntaxe abstraite (AST) de langages. Par exemple le langage OWL a plusieurs syntaxes dont une syntaxe fonctionnelle et une syntaxe RDF, de même que le langage RIF (Rule Interchange Format) ; SPIN est une notation qui permet de représenter des requêtes SPARQL en RDF.

Cet article traite du problème de la transformation d'un arbre abstrait RDF d'un langage dans sa syntaxe concrète (appelée *pretty print*). Nous proposons une approche générique pour écrire des pretty printers basés sur SPARQL pour des AST RDF. Nous définissons un pretty printer comme un ensemble de règles de transformation traitées par un moteur de pretty print. Nous proposons une extension syntaxique de SPARQL, appelée SPARQL Template, pour faciliter l'écriture des règles de transformation et l'implémentation du moteur de transformation. Nous montrons la faisabilité de notre approche en présentant deux exemples de pretty printers opérationnels pour les langages OWL et SPIN.

Mots-clés : RDF Pretty Printing, RDF AST, SPARQL Template

1 Introduction

RDF est un langage de représentation de connaissances basé sur des graphes étiquetés, conçu pour le Web sémantique et le Web des données. En tant que langage d'échange pivot, il peut être utilisé pour représenter des arbres de syntaxe abstraite (AST) de langages.

1.1 AST RDF

Un AST est le résultat de l'analyse syntaxique (parsing) d'un texte ou d'un programme dans un langage donné, depuis sa syntaxe concrète vers un arbre de syntaxe abstraite.

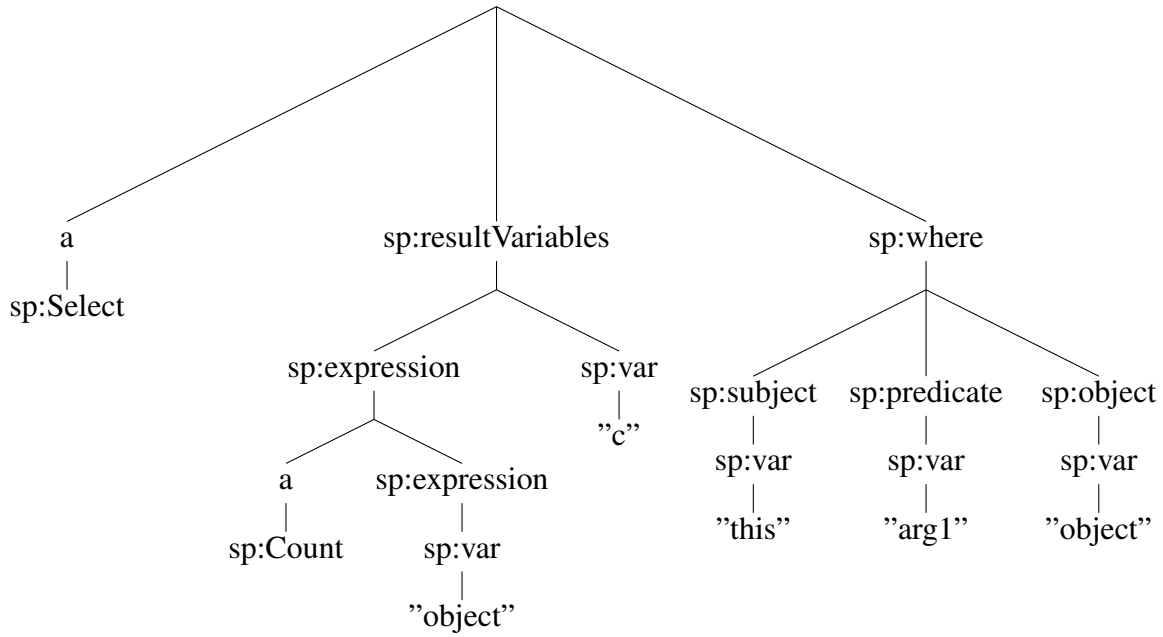
Les standards OWL et RIF sont des langages bien connus du web sémantique, munis d'une syntaxe RDF : *OWL 2 Mapping to RDF*, Patel-Schneider & Motik (2012) et *RIF in RDF*, Hawke & Polleres (2012). Un autre exemple d'AST en RDF est le format SPIN, Knublauch (2011), qui permet de représenter des requêtes SPARQL en RDF. Voici un exemple issu du document "W3C Member submission SPIN" :

```
SELECT (COUNT(?object) as ?c)
WHERE {
  ?this ?arg1 ?object
}
```

La requête SPARQL précédente est représentée par le graphe SPIN RDF suivant, où les lignes (03-09) représentent la clause SELECT et les lignes (10-14) la clause WHERE :

```
(01) @prefix sp: <http://spinrdf.org/sp#> .
(02) [] a  sp:Select ;
(03)       sp:resultVariables ([
(04)         sp:expression
(05)           [ a sp:Count ;
(06)             sp:expression [ sp:varName "object" ]
(07)           ] ;
(08)         sp:varName "c"
(09)       ]) ;
(10)      sp:where ([
(11)         sp:subject   [ sp:varName "this" ] ;
(12)         sp:predicate [ sp:varName "arg1" ] ;
(13)         sp:object    [ sp:varName "object" ]
(14)       ]) .
```

La requête SPIN précédente représente l'AST suivant :



Un AST en RDF est constitué de sommets et d'arcs où les sommets sont des ressources (URI), des blank nodes (ressources anonymes représentant des variables existentielles) ou des littéraux (des constantes). Les arcs sont des triplets RDF étiquetés par des noms de propriétés. En SPIN, les sommets intermédiaires sont des blank nodes et les feuilles sont des URI et des littéraux.
Remarquons que dans un graphe RDF les arcs d'un sommet ne sont pas ordonnés alors que dans un AST ils peuvent l'être. Quand un ordre est nécessaire, on peut utiliser une liste RDF ou

bien le Container RDF rdf:Seq. SPIN utilise une convention de nommage ad hoc : `sp:arg1`, `sp:arg2`, etc. pour ordonner les arguments.

Nous constatons donc qu'il existe un certain nombre de standards ayant une représentation sous forme d'AST RDF tels que OWL, SPIN ou RIF. Etant donné le caractère standard de RDF pour l'échange d'information, nous pouvons imaginer qu'il existe (et qu'il existera) d'autres langages ayant un AST en RDF. Ce principe de représentation d'AST RDF peut être généralisé à n'importe quel langage. Par exemple, Follenfant *et al.* (2012) proposent une syntaxe RDF pour des requêtes SQL. Egalement, nous avons prototypé un AST RDF pour des expressions mathématiques (en notation préfixe) dont la représentation pourrait ainsi être embarquée dans l'annotation RDF d'un article scientifique par exemple.

Si l'on considère le fait que RDF peut servir à décrire des énoncés de langages sous forme d'AST, on constate immédiatement que l'on peut interroger ces énoncés à l'aide de requêtes SPARQL. Par exemple : trouver les requêtes SPIN opérant sur la propriété `foaf:name`, rechercher les classes OWL sous-classes de `foaf:Human`, etc. Se pose alors le problème de la présentation des résultats des requêtes SPARQL. En effet, une requête SPIN est représentée sous forme d'un blank node RDF, de même pour les classes OWL définies et le résultat d'une requête SPARQL recherchant une requête SPIN ou une classe OWL est donc le blank node les représentant. Pour améliorer l'interaction avec l'utilisateur il est souhaitable de produire plutôt les requêtes SPIN ou les classes OWL résultats dans la syntaxe concrète de ces langages, c'est à dire dans la syntaxe concrète de SPARQL ou la syntaxe fonctionnelle de OWL.

Ainsi la possibilité de représenter des énoncés de langage sous forme d'AST RDF mène naturellement à la question de la présentation de tels énoncés dans leur syntaxe concrète, et donc de la possibilité de pouvoir réaliser une opération appelée *pretty printing* en théorie de langages. En généralisant cette question, nous posons le problème de la présentation de graphe RDF dans un format compréhensible par un utilisateur. Un autre scénario relatif à la présentation de résultats de requêtes SPARQL est par exemple la production de pages HTML issues de descriptions RDF, par exemple pour réaliser une navigation hypertexte dans une base RDF, telle que DBpedia.

1.2 Langages de Pretty Printing

Comme nous l'avons dit précédemment, un AST est le résultat d'une analyse syntaxique (parsing). Le problème inverse du *parsing*, appelé *pretty printing*, consiste à engendrer la syntaxe concrète d'un énoncé d'un langage à partir de son AST. Des langages spécifiques ont été conçus pour résoudre ce problème comme PPML, Théry (2003), pour les langages de programmation et XSLT, Kay (2007), pour XML. Ces langages reposent sur des règles de transformation déclaratives qui s'appliquent aux différents énoncés du langage cible.

PPML a été conçu dans le cadre du générateur d'environnement de programmation *Centaur*, Borras *et al.* (1988). Une règle a une partie gauche qui est une description du sous-arbre à pretty printer, et une partie droite qui est la spécification de la présentation à engendrer. L'exemple suivant montre une règle pour un énoncé addition :

```
plus(*x, *y) -> [<h> *x "+" *y];
```

XSLT a été conçu pour transformer et pretty printer des arbres XML. La même règle pour l'addition est montrée ci-dessous :

```
<xsl:template match='plus'>
  <xsl:apply-templates select='x' />
  <xsl:text> + </xsl:text>
  <xsl:apply-templates select='y' />
</xsl:template>
```

Plus particulièrement, nous avons identifié des langages de pretty print pour le Web sémantique. Le plus connu est Fresnel, Bizer *et al.* (2005), conçu pour engendrer des formats de présentation pour RDF. Il est muni d'un vocabulaire RDF permettant de définir des formats de présentation pour des types de ressources RDF et de spécifier quelles propriétés doivent être affichées et comment. Voici un exemple d'énoncé Fresnel qui indique les propriétés à afficher pour une ressource de type `foaf:Person` :

```
PersonLens a fresnel:Lens ;
  fresnel:classLensDomain foaf:Person ;
  fresnel:showProperties (
  foaf:name
  foaf:mbox
  foaf:depiction
) .
```

Voici un format de présentation pour la propriété `foaf:name` :

```
:nameFormat a fresnel:Format ;
  fresnel:label "Name" ;
  fresnel:propertyFormatDomain foaf:name .
```

Xenon RDF Stylesheet, Quan (2005), est un langage inspiré de XSLT, conçu pour appliquer des lentilles (*lenses*) sur des données RDF. La syntaxe de Xenon est RDF/Turtle utilisé comme un AST. Un énoncé de base du langage est ainsi un triplet RDF avec la propriété `xe:applyTemplates`. Le langage SPARQL est utilisé pour sélectionner les templates et les ressources dans le graphe à afficher. Le format du résultat est XHTML.

OWL-PL, Brophy & Heflin (2009), est une adaptation et une extension de XSLT pour transformer RDF/OWL en XHTML. Il vise à adapter le traitement des arbres XML aux graphes RDF. En particulier il permet d'apparier des propriétés de ressources au lieu de nœuds XML.

Alkhateeb & Laborie (2008) proposent une extension du langage SPARQL pour engendrer un document XML en réponse à une requête. Une requête SPARQL est complétée avec un template XML qui référence des variables SPARQL. Ces variables sont liées à des valeurs solutions par l'appariement d'une clause WHERE standard et le template est complété avec les valeurs de ces variables. C'est une requête de type CONSTRUCT où le template est en XML au lieu d'être en RDF.

1.3 Problèmes de recherche posés

Dans cet article, nous répondons aux problèmes de recherche suivants :

1. Comment engendrer la syntaxe concrète d'un langage depuis un AST en RDF, par exemple pour engendrer des énoncés OWL en syntaxe fonctionnelle à partir de la représentation en RDF d'une ontologie OWL.

2. Comment proposer une solution *générique* au pretty-print d'AST RDF, indépendante du langage traité et donc de la syntaxe concrète à engendrer.

3. Comment utiliser SPARQL pour écrire des règles de pretty printing.

2 SPARQL comme langage de Pretty Printing

De manière générale, une règle de pretty print comporte une condition qui s'apparie à un énoncé et une présentation qui définit le résultat du pretty print de l'énoncé filtré par la condition. Le langage SPARQL est un bon candidat pour un langage de pretty-print : dans une requête de la forme SELECT, la clause WHERE joue le rôle de la condition qui sélectionne un sommet de l'AST et la clause SELECT retourne un résultat. Considérons par exemple l'énoncé OWL suivant représentant une *restriction* de propriété dans la syntaxe RDF :

```
[
 a owl:Restriction ;
   owl:onProperty ex:hasFather ;
   owl:allValuesFrom ex:Man
]
```

La requête ci-dessous sélectionne de telles *restrictions* OWL :

```
(01) SELECT ?p ?c
(02) WHERE {
(03)  ?in a owl:Restriction ;
(04)    owl:onProperty ?p ;
(05)    owl:allValuesFrom ?c .
(06) }
```

La clause SELECT retourne les valeurs associées aux variables, mais pas (encore) le pretty print de la *restriction* en syntaxe fonctionnelle.

Supposons maintenant que nous disposons d'un ensemble de règles de pretty print, telles que celle ci-dessus, pour tous les énoncés du langage considéré, ici OWL. Supposons également que nous disposons d'une fonction `st:apply-templates` qui, appelée sur un sommet, retourne le résultat de son pretty print par les règles de présentation appropriées. Nous pouvons alors compléter la règle comme suit :

```
(01) SELECT {
(02)  (st:apply-templates(?p) as ?pp)
(03)  (st:apply-templates(?c) as ?pc)
(04) }
(05) WHERE {
(06)  ?in a owl:Restriction ;
(07)    owl:onProperty ?p ;
(08)    owl:allValuesFrom ?c .
(09) }
```

Nous avons alors (quasiment) résolu le problème. `?pp` est le résultat du pretty-print de `onProperty` et `?pc` de `allValuesFrom`. Il reste simplement à compléter par le pretty-print de la restriction. Par convention le résultat final est lié à la variable `?out`, le sommet courant étant `?in`. Ainsi, la requête finale permettant le pretty-print de la restriction est la suivante :

```
(01) SELECT {
(02)  (st:apply-templates(?p) as ?pp)
(03)  (st:apply-templates(?c) as ?pc)
(04)  (concat("ObjectAllValuesFrom(",?pp," ",?pc,")") as ?out)
(05) }
(06) WHERE {
(07)  ?in a owl:Restriction ;
(08)    owl:onProperty ?p ;
(09)    owl:allValuesFrom ?c .
(10) }
```

SPARQL permet donc d'écrire des règles de pretty-print, modulo l'écriture de la fonction d'extension `st:apply-templates` que nous allons détailler dans la suite de l'article.

3 Extension de SPARQL pour représenter des templates de transformation

Pour simplifier l'écriture des règles de pretty print, nous proposons une extension de la syntaxe du langage SPARQL, appelée SPARQL Template. La partie WHERE est standard et la partie TEMPLATE permet d'écrire directement le format de présentation à engendrer. Le template ci-dessous correspond à la requête ci-dessus.

```
(01) TEMPLATE {
(02)   "ObjectAllValuesFrom(" ?p " " ?c ")"
(03) }
(04) WHERE {
(05)  ?in a owl:Restriction ;
(06)    owl:onProperty ?p ;
(07)    owl:allValuesFrom ?c .
(08) }
```

Les règles de pretty print se focalisent sur un sommet de l'arbre pour lequel il s'agit de retourner un format de présentation. On appelle *focus node* le sommet courant qui fait l'objet d'un calcul de présentation à un instant donné du parcours de l'arbre. Le focus node est matérialisé dans une règle de pretty print par la variable `?in`, c'est une convention syntaxique car la notion de focus node n'existe pas en SPARQL. A partir d'un focus node `?in` dans la partie WHERE, le pretty printer parcourt l'arbre dans son voisinage à la recherche de ses voisins d_i. Toute mention d'un voisin d_i dans la partie TEMPLATE fait de celui-ci un nouveau focus node du moteur de pretty print, et cela récursivement jusqu'aux feuilles de l'arbre. Une occurrence d'un d_i dans la partie TEMPLATE est interprétée comme : *insérer ici le résultat du pretty print de* d_i.

Le résultat retourné par un template est par convention la valeur liée à la variable `?out` ; c'est le résultat retourné par la fonction `st:apply-templates` appliquée à cette variable. Etant donnée une liste de templates, la fonction `st:apply-templates` évalue successivement les templates jusqu'à ce qu'un template réussisse sur le focus node, c'est-à-dire retourne un résultat. Si aucun template ne réussit, le sommet lui-même est retourné. Dans le cas où la partie WHERE retourne plusieurs solutions, la partie TEMPLATE est appliquée autant de fois qu'il y a de solutions et le résultat final est la concaténation des résultats élémentaires.

L'évaluateur SPARQL est appelé avec une liaison dynamique de la variable IN (`?in`) qui est liée au focus node. Les appels récursifs à la fonction `st:apply-templates` réalisent le parcours récursif de l'AST sur les focus nodes. Le pseudo-code ci-dessous donne un aperçu du principe de la fonction `st:apply-templates`.

```
(01) Node st:apply-templates(Node node){
(02)   for (Query q : getTemplates()){
(03)     Mappings map = eval(q, IN, node);
(04)     Node res = map.getResult(OUT);
(05)     if (res != null) return res;
(06)   }
(07)   return node;
(08) }
```

Outre ce code, le pretty printer teste et gère les éventuels cycles dans le cas où le graphe RDF a des cycles (cas général). Une pile garde la trace des templates appliqués et des focus nodes de telle sorte que l'on n'applique pas deux fois le même template sur le même sommet.

3.1 Syntaxe

Nous présentons ici la syntaxe de l'extension TEMPLATE dans la syntaxe de SPARQL 1.1 Query Language, Harris & Seaborne (2013).
`Prologue` définit les préfixes, `PrimaryExpression` est une constante, une variable, un appel de fonction ou une expression parenthésée.

```
Template ::= Prologue
             'template' (iri) ? '{'
                  ( PrimaryExpression | Group ) *
                  ( Separator ) ? '}'
             WhereClause
             SolutionModifier
             ValuesClause

Group ::= 'group' ( 'distinct' ) ? '{'
             PrimaryExpression *
             ( Separator ) ? '}'

Separator ::= ';' 'separator' '=' String
```

Le namespace du langage est le suivant :

```
prefix st: <http://ns.inria.fr/sparql-template/>
```

3.2 Compilation

Nous avons développé un compilateur qui traduit les templates en requêtes SPARQL standard SELECT-WHERE. Nous présentons ici le schéma de compilation d'un template sous forme d'une requête SELECT-WHERE, au moyen d'une fonction de traduction `tr`. Cette fonction remplace les variables `V` par `coalesce(st:apply-templates(V), "")` et concatène les résultats avec `concat()`.

```
(01) tr(template(List l)) -> select (concat(tr(l)) as ?out)
(02) tr(List l) -> List(tr(ei)) pour ei élément de l
(03) tr(Literal d) -> d
(04) tr(Variable v) -> coalesce(st:apply-templates(v), "")
(05) tr(group(List l)) -> group_concat(tr(l))
(06) tr(Exp f) -> f
```

La clause `WHERE` et tout ce qui la suit sont laissés tels quels.

Soit Q la requête SPARQL SELECT-WHERE résultat de la compilation d'un template T. Soit M le multiset de solutions résultat de l'évaluation de Q. Le résultat de l'évaluation de T est calculé en appliquant un aggrégat `group_concat` sur la variable `?out` sur le multiset M :

```
Aggregation((?out), group_concat, separator, M)
```

3.3 Fonctions de pretty print

Les appels de fonction sont autorisés dans la clause TEMPLATE, comme dans une requête SPARQL standard, comme par exemple `xsd:string(?x)`. Nous avons défini un ensemble de fonctions de "pretty print" :

- `st:turtle(term)` renvoie un terme RDF au format Turtle.
- `st:uri(term)` renvoie `st:turtle(term)` si l'argument est un URI ; sinon renvoie `st:apply-templates(term)`.
- `st:call-template(name, term)` exécute un template nommé sur un focus node. Elle s'apparente à la fonction XSL `xsl:call-template`.
- `st:call-templates-with(pp, name, term)` similaire à la fonction précédente en précisant un pretty printer.
- `st:apply-templates(term)` appelle le pretty printer sur un focus node. Elle s'apparente à la fonction XSL `xsl:apply-templates`.
- `st:apply-templates-with(pp, term)` similaire à la fonction précédente en précisant un pretty printer.

- `st:apply-all-templates(term)` appelle le pretty printer sur un focus node et exécute tous les templates qui s'appliquent ; renvoie la concaténation des résultats des différents templates. Il est possible de spécifier un séparateur :
  `st:apply-all-templates(?x ; separator = ", ").`

Nous avons étendu l'interprète SPARQL de telle manière qu'il puisse évaluer *directement* une requête SPARQL de la forme TEMPLATE (en plus des cinq formes SELECT, CONSTRUCT, DESCRIBE, ASK, UPDATE).

3.4 Patrons de conception

Si la base de templates considérée contient un template nommé `st:start`, il est utilisé comme template de départ par la fonction `st:apply-templates-with`. Dans le cas contraire, le premier template qui réussit est le point de départ.

Les templates nommés sont appelés explicitement avec la fonction `st:call-template`. Il est possible d'exécuter tous les templates s'appliquant sur un focus node avec la fonction `st:apply-all-templates()` :

```
(01) TEMPLATE {
(02)    st:apply-all-templates(?x ; separator = "\n")
(03) }
(04) WHERE {
(05)   ...
(06) }
```

Il est possible d'effectuer des calculs en utilisant des appels récursifs sur des templates nommés. Il est par exemple possible d'engendrer le développement de la fonction factorielle avec le template suivant :

```
(01 TEMPLATE st:rec {
(02)  if (?i = 1, 1,
(03)    concat(?i, " * ", st:call-template(st:rec, ?i - 1)))
(04) }
(05) WHERE {
(06)    bind(?in as ?i)
(07) }
```

Le résultat de l'exécution de `st:call-template(st:rec, 3)` est : `3 * 2 * 1`

4 Applications et validation

SPARQL Template est implémenté dans la plate-forme Web sémantique Corese[1], Corby *et al.* (2012); Corby & Faron-Zucker (2010). Cinq pretty printers sont actuellement définis, pour les

[1] `http://wimmics.inria.fr/corese`

langages OWL, SPIN, SQL, Turtle et HTML, ainsi qu'un pretty printer d'expressions mathématiques en RDF vers Latex. Ces pretty printers sont accessibles en ligne[2].

Le pretty printer SPIN a été testé avec succès sur la base de tests du W3C pour SPARQL 1.1 Query & Update[3] (444 requêtes testées). Le pretty printer OWL a été testé avec succès sur l'ontology du W3C OWL 2 Primer[4], avec un chargement de la syntaxe fonctionnelle engendrée dans Protégé pour validation.

5 Discussions

5.1 Liaison dynamique de variables

Pour exécuter des templates, l'interprète SPARQL doit être capable de traiter les liaisons dynamiques de variables pour le focus node. En principe, cela pourrait être fait en SPARQL 1.1 en utilisant la clause `VALUES` qui est faite pour cela. Mais cette clause ne permet pas de passer des blank nodes en argument. Or les AST reposent en grande partie sur les blank nodes. Il est possible de réaliser le passage dynamique d'argument en utilisant une fonction d'extension dans un énoncé `bind` dans la partie WHERE comme le montre l'exemple suivant :

```
TEMPLATE { ... }
WHERE {
   bind(st:getFocusNode() as ?in)
   ...
}
```

5.2 Ordre de priorité sur les templates

Il peut être nécessaire de spécifier une priorité sur les templates pour pouvoir les considérer dans un certain ordre. Pour cela, nous utilisons une clause PRAGMA non standard avec laquelle nous avons étendu le langage SPARQL, dans laquelle nous exprimons une priorité avec une propriété `st:priority`:

```
TEMPLATE { }
WHERE { }
PRAGMA { st:template st:priority 1 }
```

5.3 Templates et régimes d'inférence

Notre interprète SPARQL implémente pour les requêtes de la forme TEMPLATE les mêmes régimes d'inférence que pour les formes standards de requêtes SPARQL (SPARQL 1.1 Entailment Regimes, Glimm & Ogbuji (2013)). Un AST peut donc être typé par une ontologie du langage cible (e.g. des classes d'énoncés) et les templates peuvent exploiter les inférences issues de cette ontologie.

[2] `ftp://ftp-sop.inria.fr/wimmics/soft/pprint`

[3] `http://www.w3.org/2009/sparql/docs/tests/data-sparql11/`

[4] `http://www.w3.org/TR/owl2-primer/`

5.4 Usages particuliers de pretty printers

Etant donné un langage muni d'une syntaxe RDF, SPARQL peut être utilisé pour rechercher des énoncés dans un tel langage, comme par exemple interroger une ontologie OWL pour rechercher les classes OWL reliées à la classe `foaf:Person`. Notre pretty printer permet alors de retourner à l'utilisateur les résultats dans la syntaxe concrète du langage cible, dans le cas de OWL, dans sa syntaxe fonctionnelle. Ceci ouvre des perspectives en termes d'interactions homme-machine, comme le montre l'exemple ci-dessous à la ligne (02) :

```
(01) SELECT ?x
(02)   (st:apply-templates-with(st:owl, ?x) as ?t)
(03) WHERE {
(04)   ?x a owl:Class ;
(05)      rdfs:subClassOf* foaf:Person
(06) }
```

Le pretty printer peut également être utilisé dans un filtre pour faire une recherche plein texte sur des énoncés dans la syntaxe concrète de leur langage.

```
(01) SELECT * WHERE {
(02)   ?x a owl:Class
(03)   FILTER(contains(
(04)      st:apply-templates-with(st:owl, ?x),
(05)      "intersectionOf"))
(06) }
```

6 Conclusion

Nous avons présenté une méthode générale pour écrire des pretty printers pour des arbres de syntaxe abstraite écrits en RDF. La méthode s'applique également à des graphes RDF quelconques, c'est-à-dire qu'il est possible d'écrire un pretty printer dédié à un domaine de connaissance quelconque modélisé en RDF. Un pretty printer est un ensemble de règles de transformation écrites dans une extension de SPARQL appelée SPARQL Template.

Nous avons validé cette méthode en réalisant des pretty printers opérationnels pour SPIN, OWL, SQL Turtle. La méthode est implémentée dans la plate-forme Web sémantique Corese. Un pretty printer peut être écrit pour engendrer la syntaxe concrète d'un langage quelconque (pourvu que ce langage soit muni d'une syntaxe RDF), mais aussi des énoncés en langue naturelle à partir de données RDF, du code HTML, etc. Nous travaillons actuellement au développement d'un pretty-printer permettant d'engendrer une représentation en Latex d'expressions mathématiques contenues dans des données RDF.

Remerciements

Nous remercions Abdoul Macina et Corentin Follenfant pour leur participation à l'écriture du pretty printer pour SQL.

References

ALKHATEEB F. & LABORIE S. (2008). Towards Extending and Using SPARQL for Modular Document Generation. In *Proc. of the 8th ACM Symposium on Document Engineering*, p. 164–172, Sao Paulo, Brésil: ACM Press.

BIZER C., LEE R. & PIETRIGA E. (2005). Fresnel - A Browser-Independent Presentation Vocabulary for RDF. In *Second International Workshop on Interaction Design and the Semantic Web @ ISWC'05*, Galway, Ireland.

BORRAS P., CLÉMENT D., DESPEYROUX T., INCERPI J., KAHN G., LANG B. & PASCUAL V. (1988). Centaur: the system. In *Proc. SIGSOFT, 3rd Annual Symposium on Software Development Environments*, Boston, USA.

BROPHY M. & HEFLIN J. (2009). *OWL-PL: A Presentation Language for Displaying Semantic Data on the Web*. Technical report, Department of Computer Science and Engineering, Lehigh University.

CORBY O. & FARON-ZUCKER C. (2010). The KGRAM Abstract Machine for Knowledge Graph Querying. In *IEEE/WIC/ACM International Conference*, Toronto, Canada.

CORBY O., GAIGNARD A., FARON-ZUCKER C. & MONTAGNAT J. (2012). KGRAM Versatile Data Graphs Querying and Inference Engine. In *Proc. IEEE/WIC/ACM International Conference on Web Intelligence*, Macau.

FOLLENFANT C., CORBY O., GANDON F. & TRASTOUR D. (2012). RDF Modelling and SPARQL Processing of SQL Abstract Syntax Trees. In *Programming the Semantic Web, ISWC Workshop*, Boston, USA.

GLIMM B. & OGBUJI C. (2013). *SPARQL 1.1 Entailment Regimes*. W3C Recommendation, W3C. http://www.w3.org/TR/sparql11-entailment/.

HARRIS S. & SEABORNE A. (2013). *SPARQL 1.1 Query Language*. Recommendation, W3C. http://www.w3.org/TR/sparql11-query/.

HAWKE S. & POLLERES A. (2012). *RIF In RDF*. Working Group Note, W3C. http://www.w3.org/TR/rif-in-rdf/.

KAY M. (2007). *XSL Transformations (XSLT) Version 2.0*. Recommendation, W3C. http://www.w3.org/TR/xslt20/.

KNUBLAUCH H. (2011). *SPIN - SPARQL Syntax*. Member Submission, W3C. http://www.w3.org/Submission/2011/SUBM-spin-sparql-20110222/.

PATEL-SCHNEIDER P. & MOTIK B. (2012). *OWL 2 Web Ontology Language Mapping to RDF Graphs (Second Edition)*. Recommendation, W3C. http://www.w3.org/TR/owl-mapping-to-rdf/.

QUAN D. (2005). Xenon: An RDF Stylesheet Ontology. In *Proc. WWW*.

THÉRY L. (2003). *A Table-Driven Compiler for Pretty Printing Specifications*. Technical Report RT 0288, Inria. http://hal.inria.fr/docs/00/06/98/91/PDF/RT-0288.pdf.

Définition de la sémantique des clés dans le Web sémantique : un point de vue théorique

Michel Chein[1], Madalina Croitoru[1], Michel Leclere[1], Nathalie Pernelle[2], Fatiha Saïs[2], Danai Symeonidou[2]

[1] UNIVERSITÉ MONTPELLIER 2 Montpellier, France
chein@lirmm.fr, croitoru@lirmm.fr, leclere@lirmm.fr

[2] UNIVERSITÉ PARIS SUD Paris, France
pernelle@lri.fr, sais@lri.fr, symeonidou@lri.fr

Résumé : De nombreuses approches ont été définies pour permettre le liage automatique de sources de données RDF publiées sur le Web. Certaines de ces approches sont basées sur la sélection des plus petits ensembles de propriétés pertinentes pour comparer deux données. Ces ensembles forment des clés et cette notion est similaire aux clés définies pour les bases de données relationnelles. Dans cet article, nous proposons d'explorer différentes sémantiques de clés qui peuvent être utilisées dans le cadre du Web sémantique.

1 Introduction

De nombreuses approches ont été définies pour permettre le liage automatique de sources de données RDF publiées sur le web (voir [2] pour un état de l'art). La plupart de ces approches exploitent des règles de liage qui spécifient les conditions que doivent remplir les descriptions de deux entités pour que celles-ci soient liées par un lien d'identité. Certaines de ces approches exploitent des sources de données pour apprendre des règles de liage expressives qui comportent des transformations, des mesures de similarité et des fonctions d'agrégation. Ces règles sont spécifiques au vocabulaire utilisé dans les sources de données exploitées. D'autres approches utilisent les axiomes définis dans l'ontologie tels que les clés ou les propriétés (inverse) fonctionnelles, en particulier depuis que OWL 2 permet, via le constructeur *owl :hasKey*, de déclarer qu'un ensemble de propriétés forment une clé pour une classe définie dans une ontologie. Ces clés peuvent être utilisées comme des règles logiques pour inférer des liens d'identité ou pour guider la construction de fonctions de similarité plus complexes pour lesquels des mesures de similarité élémentaires peuvent être choisies par un expert [4, 7, 9]. Certaines méthodes utilisent des clés pour réduire le nombre de paires d'instances à comparer par un outil de liage [5].

Ces clés n'étant pas toujours déclarées dans l'ontologie, certaines approches se sont intéressées à la découverte de clés à partir de données RDF. Le problème de découverte de clés à partir de sources de données RDF est similaire au problème de découverte de clés dans les bases de données relationnelles. Dans les deux cas, il s'agit d'un sous-problème de celui consistant à découvrir des dépendances fonctionnelles (DF). Une DF exprime le fait que la valeur d'un attribut est uniquement déterminée par les valeurs attribuées à un autre ensemble d'attributs. La découverte de clés à partir de données RDF diffère cependant du cadre habituel des bases de données relationnelles car, d'une part, certaines propriétés sont multivaluées et, d'autre part, certaines propriétés ne sont pas renseignées pour certaines données (i.e. valeurs nulles).

Dans [8], les auteurs découvrent des propriétés (inverse) fonctionnelles dans des sources de données où l'hypothèse du nom unique (UNA) est vérifiée. D'autres approches telles que [6, 1] découvrent des clés composées de plusieurs propriétés à partir de données RDF. Cependant, les clés découvertes n'ont pas la même sémantique. En effet, dans [6], deux données ayant au moins une valeur commune pour chaque propriété de la clé sont considérées comme référant à la même entité. Dans [1], les auteurs considèrent que deux données doivent être décrites par les mêmes ensembles de valeurs pour toutes les propriétés de la clé.

Dans cet article nous proposons de formaliser différentes notions de clés utilisables dans des applications de liage de données ou de détection de redondances dans une source de données. Plus précisement, nous définissons les notions d'interprétation et de déduction en logique du premier ordre pour ces deux sémantiques de clés. Nous montrons ensuite que pour définir l'ensemble des clés pouvant être découvertes (clés observées) dans une base de connaissances en terme de satisfiabilité, certains faits d'égalité et de différence doivent être explicités.

Après une étude de l'existant, nous définissons les deux sémantiques de clés. Puis, nous présentons leur utilisation dans un mécanisme de déduction. Nous présentons ensuite, comment nous définissons l'ensemble des clés observées pour ces deux sémantiques.

2 Etude de l'existant

La notion de clé a été introduite dans le modèle relationnel des bases de données. Pour une relation donnée, nous pouvons distinguer son schéma, l'ensemble des attributs, des ensembles de tuples qui forment les instances de la relation. Une clé est un ensemble d'attributs dont les valeurs définissent un seul tuple de la relation. Quand aucune clé n'a été définie, ou quand les clés impliquent un ensemble d'attributs trop grand ou des attributs trop complexes (e.g. une longue valeur textuelle), un nouvel attribut construit automatiquement peut être généré dont l'unicité est garantie. Dans le modèle relationnel, une clé doit être minimale i.e aucun sous-ensemble non vide de la clé ne doit être une clé.

Il est possible de représenter les instances d'un modèle relationnel en logique du premier ordre (FOL) en ordonnant les attributs de chaque relation et en traduisant chaque tuple en un atome ayant pour nom le nom de la relation et pour termes les valeurs des attributs dans l'ordre choisi. La conjonction des atomes obtenus forme alors la description de l'ensemble des instances de la relation. Cependant le contexte du web sémantique (WS) est différent du contexte des bases de données relationnelles. Tout d'abord, en base de données, un ensemble d'instances est représenté en intension par le schéma de la relation et en extension par l'ensemble des tuples. Dans le cadre du WS, un ensemble d'entités est représenté en intension par une classe (ou une expression de classe) et les instances représentées par cette classe forment seulement une partie de cette extension.

Deuxièmement, dans le cadre du WS, certaines données ne sont pas conformes à un schéma : rien n'interdit d'utiliser n'importe quel vocabulaire pour décrire les données. Il est cependant possible de déclarer le vocabulaire utilisé dans une ontologie.

Troisièmement, le WS ne fait pas l'hypothèse du monde fermé (Closed Word Assumption) qui pose que toute connaissance non démontrée est fausse. Aussi, si une instance de classe i n'est pas associée à une valeur v par la propriété p, cela n'entraîne pas que cette connaissance est fausse.

Enfin, le WS ne fait pas l'hypothèse du nom unique (UNA) qui exprime le fait que deux constantes sont sémantiquement égales si et seulement si elles sont syntaxiquement égales. Ainsi, deux instances syntaxiquement distinctes peuvent être déclarées égales ou différentes.

2.1 Définitions de clés dans le Web sémantique

Il existe dans le Web sémantique deux notions de clés : les axiomes OWL *owl :InverseFunctionalProperty* et *owl :hasKey* (*owl :InverseFunctionalProperty* existe dans la première version de OWL et `owl:hasKey` a été introduit dans OWL 2). L'axiome OWL *owl :InverseFunctionalProperty* permet de déclarer qu'une propriété est clé dans une base de connaissances RDF. Si cet axiome est déclaré pour une propriété p, alors sa sémantique logique pourrait être traduite par la règle en logique des prédicats suivante :

$$\forall x \forall y \forall z (p(x,z) \wedge p(y,z) \rightarrow x = y)$$

Dans le web sémantique, la déclaration d'un ensemble de propriétés $\{P_1, ...P_n\}$ comme étant clé pour une classe d'expression C donnée, peut être exprimée par la règle logique suivante :

$$\textbf{R1} : \forall x \forall y \forall z_1 ... z_n (C(x) \wedge C(y) \wedge \bigwedge_{i=1}^{n} (P_i(x,z_i) \wedge P_i(y,z_i)) \rightarrow x = y)$$

Cette définition est cohérente avec la sémantique de *owl :hasKey* dans OWL 2 fondée sur RDF [1]. Néanmoins, cette définition n'est cohérente, ni avec la présentation formelle de l'axiome *owl :hasKey* [2], ni avec la sémantique directe de *owl :hasKey* [3]. Dans cette définition une condition supplémentaire contraint les individus considérés à être nommés (*i.e.* ils doivent être des URIs ou des littéraux, mais ne peuvent pas être des noeuds blancs). Considérons un prédicat unaire $Const$ qui vaut $vrai$ si son argument est une constante sinon il vaut $faux$. Cette seconde sémantique de *owl :hasKey* peut être exprimée par la règle suivante :

$$\begin{aligned} \textbf{R2} : \forall x \forall y \forall z_1 ... z_n (C(x) \wedge C(y) \wedge Const(x) \wedge Const(y) \wedge \\ \bigwedge_{i=1}^{n} (P_i(x,z_i) \wedge P_i(y,z_i) \wedge Const(z_i)) \rightarrow x = y) \end{aligned}$$

Si on généralise cette notion, en gardant la cohérence avec OWL, on obtiendrait ce qui suit :

$$\begin{aligned} \textbf{R3} : \forall x \forall y \forall z_1 ... z_n (C[x] \wedge C[y] \wedge \bigwedge_{i=1}^{n} (P_i(x,z_i) \wedge P_i(y,z_i)) \\ \rightarrow x = y) \end{aligned}$$

où $C[.]$ est une formule avec exactement une variable libre qui représente une classe définie concernée par la clé.

1. Voir section 5.14 de http ://www.w3.org/TR/owl2-rdf-based-semantics/
2. Voir section 9.5 de http ://www.w3.org/TR/owl2-syntax/
3. Voir section 2.3.5 de http ://www.w3.org/TR/owl2-direct-semantics/

2.2 Validité d'une clé dans une base de connaissances

Si un ensemble de propriétés $K = \{p_1, \ldots, p_n\}$ est déclaré clé pour une classe C, alors ne pas respecter la clé K dans une source de données peut être dû à des erreurs ou à la présence de données dupliquées. Nous nommons abusivement "exceptions" les paires d'instances conduisant à des incohérences lorsque une clé K est déclarée. En se basant sur la sémantique des clés définie dans OWL2, une exception peut être vue comme une paire d'instances distinctes x et y de la classe C pour lesquelles il existe un ensemble d'instances ou de valeurs litérales $o_1, \ldots, o_n$ tel que x et y sont tous deux associés à o_i par la propriété p_i ($i = 1, \ldots, n$). On peut noter qu'il n'est pas nécessaire que x et y coïncide pour toutes les valeurs de p_i (excepté si p_i est fonctionnelle).

Cela implique, ce qui correspond à l'esprit du Web sémantique, que même si de nouvelles informations sont apprises sur x et y, il s'agira toujours d'exceptions pour K.

Néanmoins, lorsque les clés sont utilisées pour nettoyer ou lier des données, cela peut également être utile de considérer des clés pour lesquelles x et y seront des exceptions à K si toutes les valeurs de chacune des propriétés $p_i \in K$ coincident i.e s'il existe o_i tel que x est lié à o_i par p_i alors y doit aussi être lié à o_i par p_i, et vice versa. Avec une telle notion de clé, x et y peuvent cesser d'être des exceptions à K si de nouvelles connaissances sur x et y sont apprises.

2.3 Exemple Illustratif

La Figure 1 représente un exemple de graphe RDF décrivant quatre instances de la classe $Person$. La traduction de cet exemple en FOL est la suivante :

$F = Person(i_1) \wedge name(i_1, \textit{'Qié'}) \wedge firstName(i_1, \textit{'Wong'}) \wedge lab(i_1, \textit{'LRI'}) \wedge mail(i_1, \textit{'qie@ls.fr'})$
$\wedge Person(i_2) \wedge name(i_2, \textit{'Qie'}) \wedge firstName(i_2, \textit{'Tim'}) \wedge mail(i_2, \textit{'t.qie@orange.fr'})$
$\wedge lab(i_2, \textit{'LRI'}) \wedge Person(i_3) \wedge name(i_3, \textit{'Sting'}) \wedge mail(i_3, \textit{'sting@thales.fr'}) \wedge lab(i_3, \textit{'LIMSI'})$
$\wedge Person(i_4) \wedge name(i_4, \textit{'Sting'}) \wedge name(i_4, \textit{'Mills'}) \wedge lab(i_4, \textit{'LIMSI'}) \wedge mail(i_4, \textit{'st.mills@ls.fr'})$

Nous allons illustrer les différences entre deux notions de clés ainsi que les hypothèses qui peuvent être considérées lors de la découverte des clés ayant ces deux sémantiques.

Nous sommes intéressés par la dècouverte des clés pour une classe dans une base de connaissances OWL2 représentée en Logique du premier ordre (FOL).

Si l'on adopte le point de vu habituel des bases de données, on peut conclure que la propriété $\{mail\}$ est une clé observée mais que $\{lab\}$ n'est pas une clé observée mais nous ne savons pas comment considérer les deux autres propriétés $\{firstName\}$ et $\{Name\}$ à cause de la multivaluation ou de l'absence de valeurs. Une clef observée peut être vue comme un ensemble de prédicats tel que les valeurs de ces prédicats sont suffisantes pour identifier de manière unique les instances apparaissant dans la base de connaissances. Cette observation se base sur certaines hypothèses implicites. Pour conclure que $\{lab\}$ n'est pas une clef observée, nous faisons l'hypothèse implicite que i_1 et i_2 sont differents. De même, pour conclure que $\{mail\}$ est une clé, nous faisons l'hypothèse implicite que *'qie@ls.fr'* et *'t.qie@orange.fr'* sont différents. En fait, pour découvrir des clés en exploitant un fait, nous avons théoriquement besoin de détenir une connaissance complète sur les informations d'égalité ou de différence entre les termes apparais-

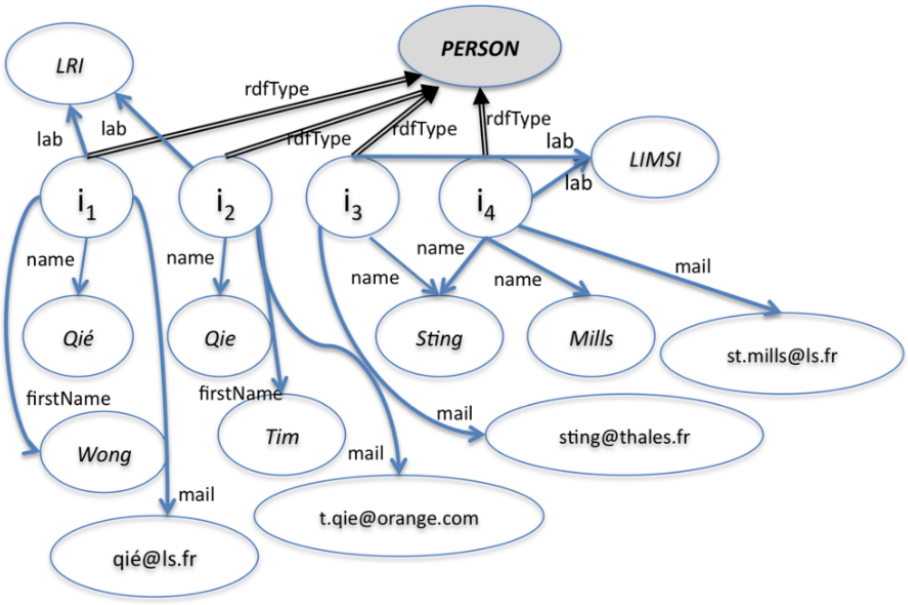


FIGURE 1 – La base de connaissances (KB_1).

sant dans le fait que ce soit pour les instances de classes ou les litéraux. Si ces informations ne sont pas connues, une heuristique simple consiste à choisir la relation d'égalité syntaxique et de ce fait à faire l'hypothèse du nom unique (UNA).

Dans certains domaines d'études, il est raisonnable de penser que pour certaines propriétés, les informations décrites pour une instance sont complètes. Cette connaissance peut découler de contraintes définies sur la propriété comme les contraintes de cardinalité ou de connaissances sur la source de données. Ainsi, on peut avoir déclaré dans l'ontologie qu'à une personne ne peut être associé qu'un laboratoire de recherche. De même, on peut savoir que dans la source de données DBLP les auteurs d'une publication sont décrits de manière exhaustive. Quand cette complétude locale est connue (i.e ce que l'on peut nommer "propriété fermée"), les clés peuvent être adaptées pour en tirer bénéfice. C'est cette intuition que nous formaliserons dans les sections suivantes sous le terme de F-règle.

Enfin, dans la sémantique que OWL2 a défini pour les clés, le prédicat "sameAs" est logiquement interprété comme une égalité. Cela n'est pas toujours pertinent sur le web de données où cette interprétation peut conduire à de nombreuses inconsistances [3]. Ainsi si l'individu "Tim Qié" de KB1, dont l'emploi principal (unique) est au laboratoire LRI, est déclaré identique à l'individu "Tim Qié" d'une autre base de connaissance KB2 dans lequel il est déclaré comme travaillant pour le LIP6, on risque en appliquant la sémantique du sameAs et l'axiome de fonctionnalité déclaré pour la propriété $\{lab\}$ d'inférer que ces deux laboratoires sont identiques. Remplacer l'égalité par une relation de similarité permet de définir une notion de clé applicable à de réels usages possibles sur le Web.

Dans la suite de cet article, nous explorons théoriquement ces différentes notions et comparons les deux notions possibles de clé.

3 Définitions de Clés

3.1 Deux définitions généralisées fondées sur les règles

L'utilisation d'une relation de similarité permet de généraliser la notion d'égalité. En effet, l'utilisation de la similarité est plus adaptée aux usages réels du Web où les similarités sont utilisées au lieu de "l'égalité logique pure".

Prenons $Sim_1, ..., Sim_n$ des relations de similarité utilisées pour comparer les valeurs de propriétés. L'introduction des relations de similarité conduit à une extension simple de la règle $R3$ (c.f. section 2.1)

Definition 1 (S-règle)
La règle de similarité ***"Some"*** *(ou* ***S-règle****) pour une classe C et pour les relations de similarité $Sim_{S_1}, \ldots, Sim_{S_n}$ est la règle définie comme suit :*

$$\forall x \forall y (C[x] \wedge C[y] \wedge$$
$$\bigwedge_{i=1}^{n} \exists z_i \exists w_i (p_i(x, z_i) \wedge p_i(y, w_i) \wedge Sim_{S_i}(z_i, w_i)) \rightarrow Sim_C(x, y))$$

Sim_{S_i} représente les relations de similarité utilisées dans la S-règle pour comparer z_i et w_i, i.e. les valeurs de la propriété p_i (e.g. mesures de similarité entre chaînes de caractères comme Levenshtein ou Jaro-Winkler). Une S-règle pour une classe C avec les relations $Sim_1, ..., Sim_n$ est notée S[(C, Sim_C), (p_1, Sim_{S_1}), ..., (p_n, Sim_{S_n})] ou, lorsqu'il n'y a pas de confusion possible, (C, Sim_C), (p_1, Sim_{S_1}), ..., (p_n, Sim_{S_n}).

Example 1
Considérons KB_1 (Figure 1). Supposons que $(Person, Sim_{Person}), (name, Sim_{S_1})$ est une S-règle et que nous avons Sim_{S_1}('Qié', 'Qie'). A partir de ces connaissances, nous pouvons inférer les similarités : $\{Sim_{Person}(i_1, i_2), Sim_{Person}(i_3, i_4)\}$

Une S-règle S est une extension d'une clé OWL dans le sens où : si tous les prédicats de similarité dans S, Sim_{S_i} et Sim_C, représentent l'égalité alors S et R_3 sont des formules équivalentes.

Proposition 1
La S-règle $S[(C, =)$; $(p_1, =)$, ..., $(p_n, =)]$ est logiquement équivalente à R_3, i.e. $R_3 \leftrightarrow S$ est valide.

Notons que pour prouver la propriété précédente, il est nécessaire de considérer la propriété de substitution pour l'égalité logique, ce qui n'est pas toujours pertinent pour `owl:sameAs` et n'est pas supposé dans cet article pour les relations de similarité.

Un deuxième type de régle de similarité peut être défini pour exprimer la seconde sémantique de clés fondée sur la comparaison des ensembles de valeurs de propriétés.

Definition 2 (F-règle)
La règle de similarité ***"Forall"*** *(ou* ***F-règle****) pour une classe C et pour les relations de similarité $Sim_{F_1}, ..., Sim_{F_n}$ est définie comme suit :*
$\forall x \forall y (C[x] \wedge C[y] \wedge \bigwedge_{i=1}^{n} (\forall z_i \exists w_i (p_i(y, z_i) \rightarrow p_i(x, w_i) \wedge Sim_{F_i}(z_i, w_i)) \wedge$
$(\forall u_i \exists v_i (p_i(x, u_i) \rightarrow p_i(y, v_i) \wedge Sim_{F_i}(u_i, v_i))) \rightarrow Sim_C(x, y)))$

Example 2
Considérons KB_1. Supposons maintenant que $(Person, Sim_{Person}), (name, Sim_{F_1})$ est une F-règle et que nous avons Sim_{F_1}('Qié', 'Qie'). A partir de ces connaissances, nous inferons la similarité : $\{Sim_{Person}(i_1, i_2)\}$

Une F-règle pour une classe C et pour les relations $Sim_{F_1}, ..., Sim_{F_n}$ est notée $F[(C, Sim_C), (p_1, Sim_{F_1}), ..., (p_n, Sim_{F_n})]$ ou simplement $(C, Sim_C), (p_1, Sim_{F_1}), ..., (p_n, Sim_{F_n})$ lorsqu'il ne peut pas y avoir de confusion.

La proposition suivante déclare que les S-règles et les F-règles sont monotones par rapport à la relation d'inclusion de propriétés.

Proposition 2
Soit R une S-règle (resp. F-règle) $[(C, Sim_C), (p_1, Sim_{R_1}), ..., (p_n, Sim_{R_n})]$ et R' une autre S-règle (resp. F-règle) $[(p_1, Sim_{R_1}), ..., (p_n, Sim_{R_n}), (p_{n+1}, Sim_{R_{n+1}}), ...]$. Alors : $R \models R'$, i.e. R' est logiquement déduite de R.

Les relations entre les S-règles et les F-règles sont étudieés dans ce qui suit.

3.2 Les modèles des S-règles et des F-règles

Dans cette section, nous présentons les propriétés des interprétations (i.e. modèles) satisfaisant les S-règles et les F-règles.

Nous montrons également que leur définition logique représente la sémantique en logique du premier ordre des S-règles et des F-règles apprises par des algorithmes existants (Propositions 3 et 4).

Des algorithmes existent dans la littérature qui permettent de découvrir des clés en s'appuyant sur la comparaison d'ensembles de valeurs. Il existe deux approches : la première [1] considère l'égalité entre ensembles de valeurs de propriétés et la seconde [6] consiste à vérifier que l'intersection entre les ensembles de valeurs de propriétés est non vide.

Definition 3
Soit $\mathcal{C}$ un ensemble, P une relation binaire sur $\mathcal{C}$, $\mathcal{S}$ une relation de similarité reflexive et symétrique définie sur le co-domaine de P, et c, c' appartenant à $\mathcal{C}$.

- *$P(c)$ indique l'ensemble d'éléments dans $\mathcal{C}$ liés à c par P, i.e. $P(c) = \{u \mid (c, u) \in P\}$;*
- *$P(c) \subseteq_s P(c')$ exprime que pour tout élément u dans $P(c)$ il existe un élément v dans $P(c')$ tel que u et v sont similaires par rapport à $\mathcal{S}$, i.e. $(u, v) \in \mathcal{S}$;*
- *$P(c) =_s P(c')$ exprime que $P(c) \subseteq_s P(c')$ et $P(c') \subseteq_s P(c)$*
- *$P(c) \cap_s P(c')$ est égale à l'ensemble $\{u \in P(c) | \exists v \in P(c')$ tel que $(u, v) \in \mathcal{S}\}$*

Example 3
Dans KB_1, nous avons par exemple, pour la relation binaire $name$ et la relation de similarité simple S (comparaison en ignorant les accents), $name(i_1) = \{$Qié$\}$, $name(i_3) \subseteq_s name(i_4), name(i_1) =_s name(i_2), name(i_2) \cap_s name(i_4) = \emptyset$.

Notons que $\cap_s$ n'est pas commutatif, néanmoins $P(c) \cap_s P(c') \neq \emptyset$ ssi $P(c') \cap_s P(c) \neq \emptyset$. Remarquons aussi que si S est la relation d'égalité alors $\subseteq_s$ est l'inclusion d'ensembles, $\cap_s$ est l'intersection d'ensembles et $=_s$ est l'égalité entre ensembles.

Definition 4
Une interprétation I en logique des prédicats dans le domaine Δ^I des prédicats apparaissant dans une S-règle, ou une F-règle, est exprimée par :
- $C^I \subseteq \Delta^I$ *l'interprétation de la classe* C *;*
- $p_i^I \subseteq \Delta^I \times \Delta^I$ *l'interprétation des predicats* p_i *;*
- $Sim_i^I \subseteq \Delta^I \times \Delta^I$ *l'interprétation des predicats* Sim_i *;*
- $Sim_C^I \subseteq \Delta^I \times \Delta^I$ *l'interépretation du predicat* Sim_C*.*

Proposition 3
Une interprétation I est un modèle pour une S-règle (C, Sim_C), (p_1, Sim_{S_1}), ..., (p_n, Sim_{S_n}) ssi quels que soient c et c' appartenant à C^I et tels que pour tout $i = 1, \ldots, k$, $p_i^I(c) \cap_S p_i^I(c') \neq \emptyset$, on a $(c, c') \in Sim_C^I$.

Notons que lorsqu'une propriété p_i^I est une fonction totale et lorsque la relation de similarité est la relation d'égalité alors une S-règle minimale (par rapport à l'inclusion de propriétés) est une clé dans le cadre des bases de données relationnelles.

Proposition 4
Une interprétation I est un modèle de F-règle (C, Sim_C), (p_1, Sim_{F_1}), ..., (p_n, Sim_{F_n}) ssi quels que soient c et c' appartenant à C^I et tels que pour tout $i = 1, \ldots, k$, $p_i^I(c) =_S p_i^I(c')$, on a $(c, c') \in Sim_C^I$.

Les relations entre les S-règles et les F-règles sont données dans la Proposition 5. La première propriété indique que la notion de S-règle est plus restrictive que la notion de F-règle, lorsque l'on considère des interpretations dans lesquelles pour tout p_i^I il existe au plus un élement de C^I n'ayant pas de valeur. Dans les bases de données relationnelles ce cas correspond au cas où il existe au plus une valeur nulle.

La deuxième propriété exprime le cas contraire lorsque l'on considère des interpretations dans lesquelles toutes les propriétés p_i^I sont fonctionnelles. Dans les bases de données relationnelles ce cas correspond au cas où toutes les propriétés sont mono-valuées.

Enfin, la dernière propriété exprime le fait que ces deux notions sont équivalentes lorsque l'interpretation de toute propriété p_i, i.e. p_i^I, est une fonction totale. En bases de données relationnelles, ce cas correspond au cas où toutes les propriétés sont renseignées et mono-valuées.

Proposition 5
Soient S-R et F-R une S-règle et une F-règle (respectivement) associées à (C, Sim_C), (p_1, Sim_{S_1}), ..., (p_n, Sim_{S_n}) .
- *Pour toute interprétation I telle que pour tout $i = 1, \ldots, n$ il existe au plus un élément $c \in C^I$ avec $p_i^I(c) = \emptyset$, on a : si I est un modèle de S-R alors I est un modèle de F-R (i.e. si $I \models S$-R alors $I \models F$-R).*
- *Pour toute interprétation I telle que pour tout $c \in C^I$ et pour toute propriété p_i on a $card(p_i^I(c)) \leq 1$, on a : si I est un modèle de F-R alors I est un modèle de S-R (i.e. si $I \models F$-R alors $I \models S$-R).*
- *Pour toute interprétation I telle que pour tout élément $c \in C^I$ et pour toute propriété p_i on a : $card(p_i^I(c)) = 1$, alors I est un modèle de S-R ssi I est un modèle de F-R (i.e. $I \models F$-R ssi $I \models S$-R).*

4 Application des S-règles et des F-règles pour la déduction de similarités

Les F-règles et des S-règles peuvent être utilisées pour inférer des similarités dans le cadre d'applications de liage de données et de nettoyage de données.

Nous présentons dans cette section l'impact sur les similarités déduites du choix de la sémantique des clés, c'est-à-dire, en utilisant les S-règles ou les F-règles.

Dans ce qui suit, on considère un *vocabulaire* V de la logique des prédicats sans symboles de fonctions et contenant au moins un prédicat unaire C, un prédicat binaire Sim_C, pour tout $i = 1, \ldots, k$ un prédicat binaire p_i, un ensemble de prédicats de similarité Sim_1,..., Sim_n réflexifs et symétriques, un ensemble de constantes et un ensemble de variables.

Definition 5 (Fait et Fait simple)

- *Un* fait relatif à une classe C *sur un vocabulaire V est la fermeture existentielle d'une conjonction d'atomes positifs construits sur V telle que les termes apparaissant dans un atome Sim_C apparaissent aussi dans les atomes de C et les termes apparaissant dans un atome Sim_{p_i} apparaissent en seconde position d'un atome p_i.*
- *Un* fait simple *est un fait sans atome Sim_C.*

Example 4
Soit F' le fait relatif à la classe C sur un vocabulaire V donné :

$F' = C(i_1) \wedge p(i_1, d) \wedge q(i_2, e) \wedge r(i_1, a) \wedge s(i_1, d) \wedge t(i_1, a) \wedge t(i_1, d) \wedge C(i_2) \wedge p(i_2, f) \wedge r(i_2, a) \wedge q(i_3, f) \wedge t(i_2, a) \wedge C(i_3) \wedge p(i_3, b) \wedge r(i_3, d) \wedge s(i_3, d) \wedge t(i_3, b) \wedge t(i_3, f) \wedge C(i_4) \wedge p(i_4, d) \wedge p(i_4, g) \wedge C(i_5) \wedge p(i_5, b) \wedge C(i_6) \wedge p(i_6, h) \wedge \exists x\ p(i_7, x) \wedge Sim_S p(f, h)$

Definition 6 (Saturation par des S-règles)
Soit $\mathcal{F}$ un fait et S une S-règle, $R(\mathcal{F})$ est le fait maximum tel que : $\mathcal{F}, S \models R(\mathcal{F})$.

Example 5
La saturation de F' en utilisant la S-règle $(C, Sim_C), (p, SimSp)$ permet d'obtenir : $R(F') = F' \wedge Sim_C(i_1, i_4) \wedge Sim_C(i_3, i_5) \wedge Sim_C(i_2, i_6)$.

L'utilisation des F-règles pour l'inférence nécessite l'enrichissement du fait $\mathcal{F}$ avec des formules exprimant la fermeture de certains prédicats (cf. [10]).

Definition 7 (Fermeture)
Soit t une instance d'une classe C et p un prédicat binaire. $Closed(t, p)$ est représenté par la formule ci-dessous qui exprime la fermeture de $\mathcal{F}$ sur p relativement à t :

$Closed(t, p) = \forall x(p(t, x) \rightarrow (x = t_1) \vee \ldots \vee (x = t_n) \vee \perp)$ avec $t_1, \ldots, t_n$ sont des termes apparaissant en seconde position de l'atome $p(t, -)$ dans $\mathcal{F}$.

$Closure(\mathcal{F}, \{p_1, \ldots, p_n\})$ est la fermeture existentielle de la conjonction d'atomes de $\mathcal{F}$ et des formules de $Closed(t, p_i)$ construite pour chaque terme t apparaissant dans un atome C et dans chaque atome p_i.

Example 6
$Closed(i_1, t) = \forall x(t(i_1, x) \rightarrow (x = d) \vee (x = a) \vee \perp)$

Definition 8 (Saturation par des F-règles)
Soit $\mathcal{F}$ un fait et F une F-règle (C, Sim_C), (p_1, Sim_{F_1}), ..., (p_n, Sim_{F_n}) , $F(\mathcal{F})$ est le fait maximal tel que :

$$Closure(\mathcal{F}, \{p_1, ..., p_n\}), F \models F(\mathcal{F})$$

Example 7
Soit r_1 une F-régle $(C, Sim_C), (p, SimSp)$. $r_1(F') = F' \wedge Sim_C(i_3, i_5) \wedge Sim_C(i_2, i_6) \wedge (Sim_C(i_7, i_1) \vee Sim_C(i_7, i_2) \vee Sim_C(i_7, i_3) \vee Sim_C(i_7, i_5) \vee Sim_C(i_7, i_6))$

Definition 9 (Deduction de similarities par une règle de similarité)
Soit $\mathcal{F}$ un fait et R une S-règle ou une F-règle. Les similarités déduites *à partir de $\mathcal{F}$ par R, notées $Sim(R, \mathcal{F})$, correspondent à l'ensemble des atomes de $R(\mathcal{F})$ n'appartenant pas à $\mathcal{F}$, i.e. $Sim(R, \mathcal{F}) = R(\mathcal{F}) \setminus \mathcal{F}$.*

Example 8
$Sim(r_1, F') = Sim_C(i_3, i_5) \wedge Sim_C(i_2, i_6)$

Notons que $Sim(R, \mathcal{F})$ produit par une règle R est réflexive et symétrique (*i.e.* Sim_C est réflexive et symétrique sur l'ensemble des termes dans $Sim(R, \mathcal{F})$). De plus, cette relation est transitive pour les F-règles puisque $=_s$ est transitive. Enfin, remarquons que la Proposition 2 conduit à ce qui suit.

Proposition 6
Soit F un fait, $R = (C, Sim_C), (p_1, Sim_{R_1}), ..., (p_n, Sim_{R_n})$.
Soit $R' = (C, Sim_C), (p_1, Sim_{R_1}), ..., (p_n, Sim_{R_n}), (p_{n+1}, Sim_{R_{n+1}})$ une S-règle (ou une F-règle). Alors, $Sim(R', \mathcal{F}) \subseteq Sim(R, \mathcal{F})$.

4.1 Découverte de clés

Definition 10 (Une clé pour un fait)
Soit $\mathcal{F}$ un fait et R une S-règle ou une F-règle : $(C, Sim_C), (p_1, Sim_{R_1}), ..., (p_n, Sim_{R_n})$. La règle R est une S ou F-clé pour C dans $\mathcal{F}$ *si $Sim(R, \mathcal{F}) \subseteq \mathcal{F}$, i.e. toute similarité déduite en appliquant R appartient au fait $\mathcal{F}$.*

Une autre distinction peut également être introduite. Selon que le fait $\mathcal{F}$ contienne ou pas des propriétés non instantiées, deux choix d'interprétation de l'absence de valeurs de propriétés peuvent être considérés : (i) considérer un modèle du contenu avec information minimale (c'est-à-dire, aucune nouvelle information ne peut être ajoutée si les valeurs sont manquantes) ou (ii) un modèle du contenu avec information non minimale (c'est-à-dire, de nouvelles informations peuvent potentiellement être ajoutées si les valeurs sont renseignées). Par abus de langage, nous notons $p_i(c) \notin \mathcal{F}$ le cas où l'instance c n'a pas de valeurs pour la propriété p_i dans le fait $\mathcal{F}$. Nous obtenons alors :

Definition 11 (Clé pour un fait avec contenu à information minimale)
Soient $\mathcal{F}$ un fait et R une S ou une F-règle : $(C, Sim_C), (p_1, Sim_{R_1}), ..., (p_n, Sim_{R_n})$. Nous Notons $\mathcal{F}'_R$ l'ensemble obtenu par la suppression de toutes les instances non complètement instantiées : $\mathcal{F}_r = \mathcal{F} \setminus \{c | \exists p_i$ tel que $p_i(c) \notin \mathcal{F}\}$. R est une S-clé ou une F-clé pour un fait avec un contenu à information minimale C dans $\mathcal{F}$ si elle est une S-clé ou une F-clé pour C dans $\mathcal{F}'_R$.

Notation : Dans ce qui suit, nous nommons ADS-clés, les F-clés pour un fait avec contenu à information minimale (puisqu'elles ont été introduites ainsi dans [1], mais d'un point de vue fonctionnel uniquement).

Supposons que $\mathcal{F}$ est complet par rapport à Sim_C, i.e. quels que soient $C(t)$ et $C(t')$ dans $\mathcal{F}$ et $Sim_C(t,t') \notin \mathcal{F}$ est interprété par le fait que t et t' ne sont pas similaires. Alors, une clé ne doit pas conduire à la déduction d'un atome $Sim_C(t,t') \notin \mathcal{F}$. Nous formalisons ceci dans ce qui suit.

Definition 12 (Fait étendu)
Un fait étendu *est composé d'un fait contenant des atomes Sim_C pouvant être niés.*

Definition 13 (Completion d'un fait)
Un fait étendu $\mathcal{F}$ relatif à une classe C est complet *si pour tout couple de termes (t,t'), non nécessairement différents, qui sont des instances de C, il contient soit $Sim_C(t,t')$ ou $\neg Sim_C(t,t')$. Soit $\mathcal{F}$ un fait, nous notons par $\mathcal{F}^C$ le* fait étendu complet *obtenu par l'ajout d'atomes de la forme $\neg Sim_C(t,t')$ pour chaque couple de termes (t,t') tels que $Sim_C(t,t') \notin \mathcal{F}$.*

Example 9
$F'^C = F' \wedge \neg Sim_C(i_1,i_2) \wedge \neg Sim_C(i_1,i_3) \wedge \ldots \wedge \neg Sim_C(i_6,i_7)$

Il est évident de noter qu'un *fait étendu $\mathcal{F}$ est consistant ssi il n'existe pas de t,t' tels que des atomes $Sim_C(t,t')$ et $\neg Sim_C(t,t')$ appartiennent tous les deux à $\mathcal{F}$.* Il est important de noter qu'un fait étendu complet est consistant.

Proposition 7
Soient $\mathcal{F}$ un fait relatif à une classe C, et S une S-règle $(C, Sim_C), (p_1, Sim_{S_1}), ..., (p_n, Sim_{S_n})$, S est une S-clé pour C dans $\mathcal{F}$ ssi $(\mathcal{F}^C, S)$ est satisfiable.

Proposition 8
Soient $\mathcal{F}$ un fait relatif à la classe C, et F une F-règle $(C, Sim_C), (p_1, Sim_{F_1}), ..., (p_n, Sim_{F_n})$, F est une F-clé pour C dans $\mathcal{F}$ ssi $closure(\mathcal{F}^C, F), F$ est satisfiable.

5 Conclusion

Dans cet article, nous avons exploré et comparé différentes sémantiques de clés que ce soit en terme d'interprétation, de déduction et de découverte. Ces clés peuvent être utilisées dans le contexte du Web de Données. Nous souhaitons, dans un premier temps compléter cette formalisation en fournissant des preuves pour les principales propositions. De plus, nous envisageons de comparer expérimentalement leur utilisation pour lier différents jeux de données de divers domaines.

Remerciements
Ce travail a été financé par l'Agence Nationale de la Recherche (projet QUALINCA-ANR-12-CORD-0012).

Références

[1] ATENCIA M., DAVID J. & SCHARFFE F. (2012). Keys and pseudo-keys detection for web datasets cleansing and interlinking. In *EKAW*, p. 144–153.

[2] FERRARA A., NIKOLOV A. & SCHARFFE F. (2011). Data linking for the semantic web. *Int. J. Semantic Web Inf. Syst.*, **7**(3), 46–76.

[3] HALPIN H., HAYES P. & THOMPSON H. S. (2011). When owl : sameas isn't the same redux : A preliminary theory of identity and inference on the semantic web. In *Proc of Workshop on Discovering Meaning On the Go in Large Heterogeneous Data*, p. 25–30.

[4] HOGAN A., ZIMMERMANN A., UMBRICH J., POLLERES A. & DECKER S. (2012). Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. *J. Web Sem.*, **10**, 76–110.

[5] ISELE R., JENTZSCH A. & BIZER C. (2011). Efficient multidimensional blocking for link discovery without losing recall. In *Proceedings of the 14th International Workshop on the Web and Databases 2011, WebDB 2011*.

[6] PERNELLE N., SAÏS F. & SYMEONIDOU D. (2013). An automatic key discovery approach for data linking. *Web Semantics : Science, Services and Agents on the World Wide Web*, **23**, 16 – 30.

[7] SAÏS F., PERNELLE N. & ROUSSET M.-C. (2009). Combining a logical and a numerical method for data reconciliation. *Journal on Data Semantics*, **12**, 66–94.

[8] SUCHANEK F. M., ABITEBOUL S. & SENELLART P. (2011). Paris : Probabilistic alignment of relations, instances, and schema. *The Proceedings of the VLDB Endowment*, **5**(3), 157–168.

[9] VOLZ J., BIZER C., GAEDKE M. & KOBILAROV G. (2009). Discovering and maintaining links on the web of data. In *Proceedings of the 8th International Semantic Web Conference*, ISWC '09, p. 650–665, Berlin, Heidelberg : Springer-Verlag.

[10] WAGNER G. (2003). Web rules need two kinds of negation. In *PPSWR*, p. 33–50.

Publication, partage et réutilisation de règles sur le Web de données

Oumy Seye[1], Catherine Faron-Zucker[1], Olivier Corby[2], Alban Gaignard[1]

[1] Univ. Nice Sophia Antipolis, CNRS, I3S, UMR 7271, 06900 Sophia Antipolis, France
{seye,faron,gaignard}@i3s.unice.fr

[2] INRIA Sophia-Antipolis Méditerranée, 06900 Sophia Antipolis, France
olivier.corby@inria.fr

Résumé : L'objectif de notre travail présenté dans cet article est de favoriser la réutilisation de règles sur le Web, basée sur les principes du Web de données. En complément de données RDF, de schémas RDFS ou d'ontologies OWL, des règles peuvent être publiées et partagées sur le Web. Notre approche consiste à considérer des bases de règles comme des sources de données, représentées en RDF, qui peuvent être publiées, partagées et interrogées sur le Web de données, permettant ainsi la sélection et la réutilisation des règles pertinentes et utiles dans un contexte ou une application particuliers. Nous envisageons la sélection de règles selon des annotations qui les décrivent ou selon leur contenu, ou les deux. Nous avons implémenté et mis en œuvre notre approche avec le moteur Corese/KGRAM permettant le traitement de données centralisées ou distribuées sur le Web de données et nous avons conduit des expérimentations sur la sélection des règles de la sémantique de OWL pour des données basées sur des ontologies populaires.

Mots-clés : Web de données liées, règles, SPARQL, RDF

1 Introduction

Le partage et la réutilisation de connaissances sont les objectifs principaux du Web de données liées. Un ensemble de modèles tels que RDF, RDFS et OWL, ont été développés pour permettre aux humains et agents logiciels de publier et accéder aux données. Les schémas RDFS et les ontologies OWL sont des standards pour représenter les connaissances d'un domaine sur le Web de données. La définition de règles d'inférence constitue un moyen complémentaire ou alternatif pour capturer la sémantique sur le Web de données. En complément de données RDF, de schémas RDFS ou d'ontologies OWL, des règles peuvent être publiées et partagées sur le Web. Dans ce contexte l'objectif principal de notre travail présenté dans cet article est de favoriser la réutilisation de règles sur le Web, en se basant sur les principes du Web de données. Dans notre approche nous considérerons des bases de règles comme des sources de données particulières qui, comme toutes sources de données, peuvent être publiées, partagées et interrogées sur le Web de données, permettant ainsi la sélection et la réutilisation des règles pertinentes et utiles dans un contexte ou une application particuliers.

Un premier scénario de réutilisation de règles publiées sur le Web est le suivant. Pour traiter une source de données RDF particulière, un utilisateur souhaite sélectionner des règles selon l'organisme qui les publie, ou bien selon leur date de publication, ou leur sujet, c'est-à-dire, plus généralement, selon des critères qui peuvent être attachés aux règles dans des méta-données, et non plus seulement selon le contenu des règles.

Un second scénario, complémentaire du précédent, est relatif à la prise en compte des connaissances de domaine lors de l'exploitation d'une source de données RDF. Pour cela, l'utilisateur met en œuvre un moteur implémentant la sémantique du schéma RDFS ou de l'ontologie OWL

associés aux données, et recherche des règles de domaine relatives à ses données. Il recherche donc sur le Web, parmi toutes les règles d'inférence publiées, celles qui sont susceptibles de s'appliquer aux données qu'il souhaite exploiter, c'est-à-dire celles dont les termes sont ceux du schéma ou de l'ontologie sur lesquels reposent ses données.

Enfin, un troisième scénario, complémentaire du précédent, est la sélection la plus fine possible des règles susceptibles de s'appliquer à une source de données considérée. Pour réduire le temps des traitements ultérieurs d'une source de données, un utilisateur cherche à identifier le sous-ensemble des classes et propriétés effectivement utilisées dans les données considérées et à réduire la base de règles considérées aux seules règles susceptibles de s'appliquer sur ses données. Par exemple, dans une application sur des sources de données de grande taille, la première étape peut être la sélection des règles de la sémantique du modèle pertinentes pour l'ontologie particulière considérée. Ainsi, dans le cas d'une ontologie qui comporterait uniquement des classes primitives, beaucoup de règles de la sémantique de RDFS ou de OWL deviennent inutiles, c'est-à-dire ne s'appliquent pas, et l'économie de la tentative de leur application devient précieuse dans le cas d'une source de données de grande taille.

Dans cette perspective, pour répondre à de tels scénarios, nous proposons de considérer les règles comme des données, qu'il s'agit donc de publier dans le langage RDF, le standard du Web de données. Cette publication sur le Web permet leur réutilisation, basée sur la recherche automatique de règles pertinentes pour une application donnée ou un contexte spécifique. Cette recherche repose sur l'interrogation du contenu des règles et/ou des méta-données qui peuvent leur être associées, avec des requêtes SPARQL, le standard du Web de données pour interroger des données RDF. En d'autres termes, pour répondre au problème de la publication et la réutilisation de règles sur le Web, nous l'envisageons comme un problème classique en ingénierie des connaissances de partage et de réutilisation de connaissances et nous proposons une approche basée sur (1) la représentation en RDF à la fois du contenu et de méta-données associées aux règles, (2) leur publication sur le Web de données et (3) la construction automatique de bases de règles spécifiques à un contexte ou une application particuliers basée sur l'interrogation de sources de données RDF représentant des règles à l'aide de requêtes SPARQL.

Dans la section suivante, nous présentons le langage de règles que nous adoptons et justifions notre choix par rapport aux alternatives possibles. Dans la sections 3, nous présentons notre approche de la sélection de règles pertinentes dans un contexte donné, selon leur annotation ou leur contenu. Dans la section 4 nous présentons une implémentation de scénarios de sélection de règles dans des sources de données distribuées et leur application sur des sources de données RDF distribuées, avec le moteur sémantique Corese/KGRAM.

2 Publication de règles sur le Web de données : choix des langages SPARQL et SPIN

RIF [1] (acronyme de *Rule Interchange Format*) est le format de règles recommandé par le W3C pour échanger des règles sur le Web. Cependant, RIF reste peu utilisé. Le langage SPARQL [2] recommandé par le W3C pour l'interrogation de données RDF précède RIF et est également utilisé comme langage de règles dans de nombreux travaux sur le Web sémantique.

SELECT et ASK sont les formes de requêtes SPARQL les plus connues : une requête de la

1. http://www.w3.org/TR/rif-overview/
2. http://www.w3.org/TR/rdf-SPARQL-query/

forme ASK permet de demander si un appariement existe entre le graphe requête et le graphe RDF ; une requête de la forme SELECT permet de demander les valeurs des variables indiquées dans la clause SELECT pour lesquelles la clause WHERE de la requête s'apparie avec le graphe RDF interrogé.

La forme CONSTRUCT permet de produire un nouveau graphe RDF en remplaçant les variables du graphe de la clause CONSTRUCT par les valeurs pour lesquelles le graphe requête de la clause WHERE s'apparie avec le graphe RDF interrogé. Une telle requête peut être vue comme une règle (la prémisse représentée par la clause WHERE et la conclusion par la clause CONSTRUCT) et son traitement comme l'application d'une règle en chaînage avant pour enrichir le graphe RDF. Voici par exemple la représentation en SPARQL de la règle « tout homme est mortel » :

```
CONSTRUCT {?x rdf:type bio:Mortal} WHERE {?x rdf:type bio:Human}
```

Parmi les travaux qui utilisent SPARQL comme langage de règles, dans (Angles & Gutierrez, 2008; Polleres, 2007), les auteurs établissent une correspondance entre SPARQL et Datalog (nr-Datalog$^{\neg}$: sans récursion, avec négation), le langage de requêtes et de règles des bases de données déductives.

Dans (Schenk & Staab, 2008) les auteurs utilisent SPARQL comme langage de règles pour définir de nouvelles données RDF à partir de sources de données existantes. La forme CONSTRUCT de SPARQL est utilisée dans des réseaux de graphes pour effectuer la correspondance des patrons de graphes sur des graphes d'entrée.

Le framework R2R qui permet de publier des mappings sur le Web et de traduire des données du Web vers un schéma local est basé sur la forme CONSTRUCT de SPARQL pour exprimer des transformations de données (Bizer & Schultz, 2010).

Avec SPARQL++, (Polleres *et al.*, 2007) utilisent le langage SPARQL comme un langage de règles pour exprimer des alignements entre vocabulaires RDF et proposent pour cela certaines extensions de la forme de requêtes CONSTRUCT.

Avec SPIN [3] (acronyme de SPARQL *Inferencing Notation*), H. Knublauch invite à considérer SPARQL comme un langage de règles et de contraintes (les formes CONSTRUCT, UPDATE et ASK) et propose une notation en RDF. SPIN est une member Submission au W3C [4] depuis 2011.

SPIN est le format de règles que nous avons adopté, qui permet de publier des règles SPARQL sur le Web de données. Nous avons développé dans le moteur Corese/KGRAM un parser pour traduire des règles SPARQL en SPIN et un pretty-printer pour produire des requêtes dans la syntaxe concrète de SPARQL à partir de leur représentation en SPIN. Nous stockons la représentation SPIN de chaque règle SPARQL dans un graphe RDF nommé. Ainsi, tous les énoncés relatifs à une règle sont regroupés dans un même graphe, ce qui facilite la gestion, la recherche et la récupération de règles. Voici les représentation en SPARQL et en SPIN/RDF de la règle exprimant que si quelqu'un a un parent qui a un frère, alors celui-ci est son oncle.

```
prefix ex: <http://www.example.org/humans#>
CONSTRUCT {?x ex:hasUncle ?z}
WHERE {?x ex:hasParent ?y. ?y ex:hasBrother ?z }
```

3. http://spinrdf.org/

4. http://www.w3.org/Submission/spin-sparql/

```
@prefix sp: <http://spinrdf.org/sp#> .
@prefix ex: <http://www.example.org/humans#> .
_:b1  sp:varName "y"^^xsd:string .
_:b2  sp:varName "z"^^xsd:string .
_:b3  sp:varName "x"^^xsd:string .
[] a sp:Construct ;
   sp:templates ([ sp:subject sp:varName _:b3 ;
                   sp:predicate ex:hasUncle ;
                   sp:object _:b2 ]) ;
   sp:where ([ sp:subject sp:varName _:b3 ;
               sp:predicate ex:hasParent ;
               sp:object _:b1 ]
             [ sp:subject _:b1  ;
               sp:predicate ex:hasBrother ;
               sp:object _:b2 ]) .
```

Cependant, notre approche et notre implémentation se généralisent à tout langage de règles muni d'une syntaxe RDF ou qui peut être traduit dans le langage SPARQL. Notamment, dans (Seye *et al.*, 2012), nous avons décrit un dialecte RIF qui peut être traduit en SPARQL. Nous avons implémenté ce dialecte avec le moteur sémantique Corese/KGRAM et nous avons déployé un service en ligne pour la traduction des règles RIF-SPARQL en SPARQL. Ainsi, nous sommes capables de publier en RDF des règles RIF de ce dialecte en les traduisant d'abord dans le langage SPARQL.

Une approche similaire pourrait également être adoptée pour SWRL [5] (acronyme de *Semantic Web Rule Language*), un langage de règles relativement utilisé bien que non standardisé (*member Submission* au W3C depuis 2004). Il est muni d'une syntaxe RDF qui permettrait de publier sur le Web de données les règles écrites dans ce langage.

3 Sélection de règles : interrogation en SPARQL de leurs représentations RDF

La publication de règles en RDF sur le Web de données permet leur partage et leur réutilisation. Ces *données* RDF peuvent en effet être recherchées de façon standard, avec des requêtes SPARQL pour sélectionner des règles intéressantes à réutiliser dans un contexte spécifique.

Dans (González-Moriyón *et al.*, 2012) les auteurs ont exploré la réutilisation de règles RIF et proposé l'outil RIF Assembler qui permet de sélectionner et réutiliser des règles RIF en interrogeant les méta-données de ces règles, en les traduisant en RDF selon l'interprétation commune de RIF en RDF. Cependant, en RIF, les méta-données ne sont pas obligatoires et les règles non annotées ne peuvent pas être réutilisées avec RIF Assembler. De plus, seules les méta-données sont exploitées dans cette approche, et pas le contenu des règles.

Nous proposons une approche *générale* et *unifiée* de la réutilisation de règles SPARQL représentées en RDF, basée sur leur sélection par interrogation de leur contenu aussi bien que des métadonnées associées.

5. http ://www.w3.org/Submission/SWRL/

3.1 Sélection basée sur l'interrogation des méta-données associées aux règles

Les règles peuvent être sélectionnées en interrogeant des métadonnées qui peuvent leur être associées, dès lors qu'elles sont identifiées par un URI : nous identifions chaque règle par l'URI du graphe nommé RDF contenant sa représentation SPIN et cette URI peut être décrite en RDF. Nous répondons ainsi au premier scénario présenté en introduction.

Les métadonnées de règles peuvent par exemple contenir des informations sur la source des règles, l'auteur, le titre, le sujet, etc. Ces métadonnées lient les règles avec d'autres schémas et données du Web. Dans la recommandation RIF du W3C, il est suggéré d'utiliser le Dubling Core et des propriétés des modèles RDFS et OWL pour annoter les règles. Par exemple, la requête suivante permet de rechercher des règles sur le benfluorex parmi les données qui pourraient être publiées par l'agence nationale de sécurité du médicament et des produits de santé (ANSM), dont certaines pourraient être représentées par des règles SPARQL.

```
PREFIX  sp: <http://spinrdf.org/sp#>
PREFIX  drug: <http://www.example.org/drug#>
PREFIX  dc: <http://dublincore.org/documents/dcmi-namespace/#>
SELECT DISTINCT(kg:pprintWith(pp:spin, ?x) as ?res)
WHERE { ?x a sp:Construct ;
          dc:source <https://icrepec.afssaps.fr/Public> ;
          dc:subject drug:Benfluorex } }
```

Remarquons que dans cette requête, les règles solutions sont associées à la variable `?x` et la fonction `kg:pprintWith` appliquée sur ces solutions appelle le pretty-printer de Corese pour produire la représentation dans la syntaxe concrète de SPARQL des règles solutions, à partir de leur représentation en SPIN/RDF (Corby & Faron-Zucker, 2014). Les règles solutions sont ainsi directement utilisables par tout moteur de règles SPARQL, et en particulier le moteur de règles de Corese/KGRAM.

Le modèle Open Annotation [6] semble également bien adapté pour l'annotation de règles, qui fait une distinction entre le corps (`oa:body`) et le but (`oa:target`) d'une annotation de ressource, et permet ainsi de distinguer méta-données et description de contenu. La notion de sélecteur (`oa:Selector`) permet de décrire séparément différentes parties d'une ressource et permet, dans le cas d'une règle, de distinguer des annotations portant sur sa prémisse ou sa conclusion.

3.2 Sélection basée sur l'interrogation du contenu des règles

Publier des règles SPARQL en SPIN permet de lier la représentation de leur contenu avec d'éven-tuelles méta-données et avec d'autres données du Web : en particulier, elles sont directement liées avec des données RDF(S) ou des ontologies OWL qui peuvent être utilisées lors de la sélection de règles par interrogation de leurs contenus. Il s'agit ici de répondre aux scénarios 2 et 3 présentés en introduction.

Par exemple, la requête SPARQL suivante permet de sélectionner toutes les règles dont la prémisse contient des propriétés représentant des relations familiales.

6. http ://www.openannotation.org/spec/core/

```
PREFIX sp: <http://spinrdf.org/sp#>
PREFIX ex: <http://www.example.org/humans#>
SELECT DISTINCT (kg:pprintWith(pp:spin, ?x) as ?res)
WHERE { ?x a sp:Construct
        ?x sp:where ?m
        ?m (!sp:void)+ ?s
        ?s sp:predicate ?z
        ?z rdfs:subPropertyOf* ex:hasFamilyRelationShip }
```

Plus généralement, il est intéressant de pouvoir sélectionner des règles dans le contenu desquelles apparaissent des classes ou propriétés appartenant à une ontologie donnée. En effet, étant donnée une source de données RDF, il est inutile de chercher à appliquer des règles qui n'utilisent pas le même vocabulaire : elles ne s'appliqueraient pas sur les données. Dans le cas de sources de données de grande taille, un tel filtrage des règles susceptibles de s'appliquer peut devenir crucial pour la faisabilité des inférences. La requête suivante permet de sélectionner les règles qui utilisent des termes appartenant à un schéma RDFS chargé.

```
PREFIX sp: <http://spinrdf.org/sp#>
SELECT DISTINCT (kg:pprintWith(pp:spin, ?x) as ?res)
WHERE {
   SELECT DISTINCT ?resource
   WHERE {
      { ?resource rdf:type rdfs:Class }
       UNION
      { ?resource rdf:type rdf:Property }
     ?x a sp:Construct .
     ?x sp:where ?m .
     ?m (! sp:void)+ ?s .
     ?s ?p ?resource  } }
```

Une autre sélection, plus restrictive, pourrait être implémentée par une requête recherchant les règles pour lesquelles *toutes* les classes et propriétés apparaissant dans la prémisse appartiennent à un vocabulaire donné.

De telles présélections de règles ne remplacent pas les techniques d'optimisation des raisonneurs basées sur l'analyse des dépendances telle que celle présentée dans (Baget, 2004). Il s'agit de construire pour une ontologie une base des règles pertinentes. Nous entendons par règles pertinentes celles susceptibles de s'appliquer sur les données et donc de produire de nouvelles données inférées. Il s'agit ici des règles contenant au moins un terme de l'ontologie. Lors de l'application de la base de règles ainsi créée à une source de données particulière, les techniques d'optimisation des raisonneurs basées sur l'analyse des dépendances de règles peuvent s'appliquer.

3.3 Sélection basée sur l'ajustement de la sémantique du modèle RDFS ou OWL à la sémantique du vocabulaire des données

Dans une application sur des sources de données de grande taille, une première étape de sélection des règles de la sémantique du modèle pertinentes pour l'ontologie particulière considérée peut être précieuse. Il s'agit ici d'un cas particulier du scénario 3. Par exemple dans le cas d'une ontologie ne comportant que des classes RDFS primitives, la plupart des règles de la

sémantique de RDFS ou celles de OWL ne s'appliqueront pas. Or la connaissance a priori des règles pertinentes peut être cruciale dans le cas du traitement d'une source de données de grande taille.

Nous avons implémenté la sémantique de RDFS en écrivant une base de 14 règles et la sémantique de OWL 2 RL en construisant une base de 71 règles d'inférence. Nous avons défini deux requêtes SPARQL génériques pour sélectionner les règles SPIN de la sémantique de RDFS ou de OWL 2 RL. Ces deux requêtes sélectionnent les règles dont la prémisse contient des ressources appartenant au méta-modèle de RDFS et/ou de OWL. Voici une version simplifiée d'une telle requête :

```
SELECT DISTINCT  (kg:pprintWith(pp:spin, ?r1) as ?res)
WHERE {
    # ?resource matches URIs in the OWL or RDFS meta-model
    SELECT DISTINCT ?resource
    WHERE {
      GRAPH ?g {
        ?o a owl:Ontology
        { ?resource ?p ?y} union {?x ?resource ?y} union {?x ?p ?resource}
        filter ( ?resource in (rdf:Property, rdf:type)
                    || strstarts(?resource, owl:)
                    || strstarts(?resource, rdfs:) )
      }
    }
    # ?r1 contains a resource from the above ontology
    ?r1 a sp:Construct
    ?r1 sp:where ?w1
    ?w1 (! sp:nil)+ ?x
    ?x ?p ?resource
    VALUES ?p { sp:subject sp:predicate sp:object }
}
```

La requête finale contient également les règles qui dépendent des précédentes, c'est-à-dire que l'application des précédentes peut rendre à leur tour déclenchables : la prémisse de ces dernières peut être appariée avec la conclusion de l'une des premières. La requête de sélection complète comporte 80 lignes que nous ne reproduisons pas ici.

Nous montrons dans la section suivante que la présélection de règles de la sémantique de RDFS ou de OWL pertinentes pour un schéma ou une ontologie et donc susceptibles d'être appliquées sur une source de données reposant sur ce vocabulaire peut permettre de réduire considérablement le temps d'application des règles sur les données interrogées.

3.4 Expérimentations

Nous avons construit des bases de règles SPIN/RDF implémentant la sémantique de RDFS et OWL 2 RL et nous montrons dans cette partie comment sélectionner dans ces bases les règles pertinentes pour des ontologies populaires du Web de données et quelle réduction du coût des inférences cela représente lors de l'application de ces bases de règles à une source de données basée sur une de ces ontologies.

3.4.1 Sélection des règles de la sémantique de RDFS et OWL pertinentes pour une ontologie donnée

FOAF

L'ontologie FOAF [7] permet de décrire des personnes et leurs relations. En utilisant la requête de sélection présentée dans la section 3.3, nous sélectionnons 42 règles pertinentes pour cette ontologie parmi les 71 règles de la sémantique de OWL 2 RL. Pour implémenter la sémantique de RDFS, la requête idoine sélectionne 10 règles parmi les 14 de la base complète.

DBpedia

L'ontologie DBpedia [8] est une ontologie générale pour représenter en RDF des données issues de Wikipedia. En utilisant la même requête de sélection, 40 règles de la sémantique de OWL 2 RL sont sélectionnées parmi les 71 initiales comme pertinentes pour cette ontologie. Pour implémenter la sémantique de RDFS, 6 règles sont extraites parmi les 14 règles de la base complète.

INSEE

L'ontologie géographique de l'INSEE [9] permet de décrire les données issues du Code officiel géographique (COG) concernant notamment les régions, les départements, les arrondissements, les cantons et les communes. Pour cette ontologie, la phase de sélection de règles permet de réduire l'ensemble de règles à appliquer à 52 règles de la sémantique de OWL 2 RL pertinentes parmi les 71 de la base complète.

Pour chacune de ces ontologies, l'application des règles d'inférence RDFS et OWL 2 RL, avec et sans sélection de règles, engendrera le même nombre de triplets inférés. Cependant, la phase de sélection de règles pertinentes permet de réduire significativement le temps de calcul de ces inférences. La section suivante illustre plus précisément les gains obtenus.

3.4.2 Mesure de gain de temps lors de l'application de règles pré-sélectionnées

Dans ces expérimentations, nous nous appuyons sur la base des 71 règles SPARQL que nous avons écrites pour implémenter la sémantique de OWL 2 RL, le jeu de données RDF DBpedia-person, et le moteur de règles en chainage avant de Corese/KGRAM. Nous avons mesuré le temps d'exécution de l'ensemble des règles, avec et sans sélection, sur une machine dotée de 32 Go de RAM et de deux CPUs Intel quad-core cadencés à 2.2 GHz.

Le tableau 1 illustre les temps moyens d'application des règles sur cinq exécutions. Ces résultats montrent que la sélection de règles apporte un gain de temps de plus de 22% lors de l'application des règles OWL 2 RL sur les données DBpedia-person (1,7 million de triplets).

7. `http://xmlns.com/foaf/spec`
8. `http://downloads.dbpedia.org/3.9/dbpedia_3.9.owl.bz2`
9. `http://rdf.insee.fr/def/geo/insee-geo-onto.ttl`

Données DBpedia-person	*Toutes les règles*	*Règles pertinentes*
Nombre de règles appliquées	71	40
Nombre de triplets initiaux	1745628	1745628
Nombre de triplets initiaux et inférés	3835867	3835867
Temps moyen de calcul des inférences (s)	852,3	659,5
Ecart-type (s)	6,9	15,4

TABLE 1 – Réduction du temps de calcul des inférences OWL 2 RL lorsque les règles sont pré-sélectionnées.

4 Sélection et application de règles distribuées sur des données distribuées

Nous avons mis en œuvre notre approche de publication et de réutilisation de règles en utilisant le moteur sémantique Corese/KGRAM qui permet d'interroger des sources de données distribuées et d'appliquer des règles d'inférence sur celles-ci.

4.1 Recherche sémantique en environnement distribué

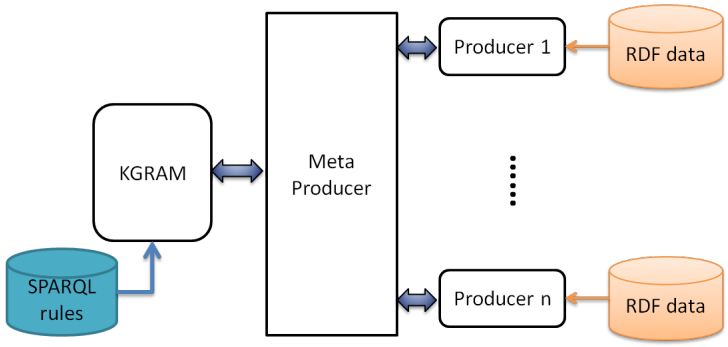
Corese/KGRAM est un moteur sémantique conçu pour l'interrogation de graphes RDF (Corby *et al.*, 2004). Corese/KGRAM implémente le langage de requête SPARQL 1.1 et permet donc d'interroger des données RDF avec des requêtes de la forme SELECT ou ASK mais aussi d'effectuer des mises à jour sur les données RDF avec des requêtes de la forme CONSTRUCT ou des opération INSERT/DELETE (SPARQL UPDATE).

Corese/KGRAM repose sur une interface `Producer` pour l'énumération de triplets RDF. Dans le cas où les données sont réparties, un `Metaproducer` itère sur les sources disponibles et énumère, en parallèle et de manière transparente, des triplets provenant de ces sources. Finalement, une interface d'invocation à distance permet la fédération de sources de données RDF distribuées sur le Web de données (Corby *et al.*, 2012).

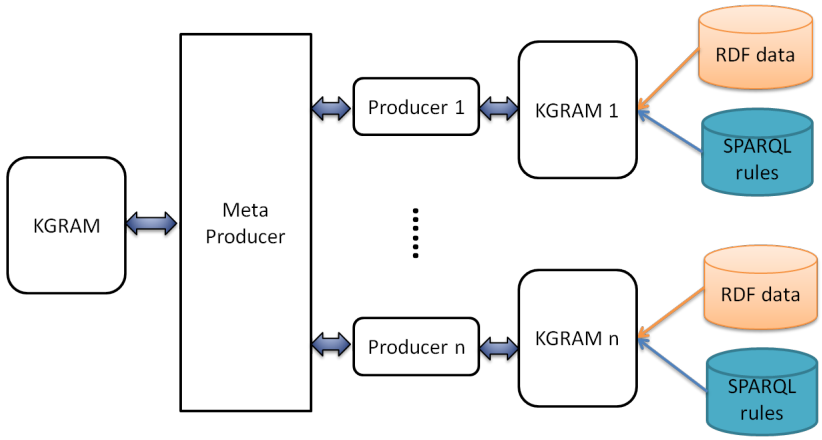
4.2 Application de règles d'inférence en environnement distribué

Avec son méta-producteur, Corese/KGRAM peut s'appuyer directement sur son moteur de règles pour raisonner sur des données multi-sources. La Figure 1 illustre deux scénarios mettant en jeu des règles et des données réparties. Le premier scénario montre l'application d'une base de règles centralisée, sur des données distribuées. Le second scénario illustre l'application de règles d'inférences elles-mêmes réparties, sur des données également réparties. Ce scénario fait sens lorsque un ensemble de règles est spécifique à un jeu de données, mais peut également s'appliquer sur d'autres sources de données.

De plus, comme nous représentons les règles en SPIN/RDF, la sélection de règles en amont de leur application peut également se faire dans un environnement distribué (scénarios 2 de la Figure 1). Des sources de données SPIN/RDF distribuées sont interrogées pour sélectionner des règles pertinentes et construire une base de règles à appliquer à des sources de données distribuées.



Scenario 1: a centralized SPARQL rule base and distributed RDF sources



Scenario 2: distributed RDF sources with associated SPARQL rule bases

FIGURE 1 – Scénarios de distribution des données et des règles

4.3 Preuve de concept

En combinant ainsi la sélection des règles représentées en SPIN/RDF et leur application à des sources de données RDF, Corese/KGRAM permet de répondre à des scénarios complexes du Web de données, intégrant la sélection et l'application de règles sur des données du Web.

Nous avons mené des premières expériences sur les données du tutoriel de Corese/KGRAM. Nous avons traduit (en utilisant notre pretty printer SPIN) la base des 25 règles SPARQL du tutoriel en un ensemble de 25 graphes SPIN/RDF que nous avons mis en ligne. Nous avons divisé les données RDF en deux sources de données RDF que nous avons mises en ligne : `human1.rdf` et `human2.rdf`. Nous avons sélectionné un *sous-ensemble* du schéma RDFS `human.rdfs` que nous avons mis en ligne dans une source de données RDFS supplémentaire `test.rdfs`.

Nous avons alors utilisé Corese/KGRAM pour interroger les 25 sources données SPIN/RDF et sélectionner les seules règles dont l'hypothèse fait référence à des ressources décrites dans le schéma `test.rdfs`.

Enfin, nous avons centralisé les 7 règles ainsi sélectionnées, étant donné le sous-ensemble de l'ontologie où nous nous limitons, et nous avons utilisé le moteur de règles de Corese/KGRAM pour appliquer ces règles sur les données RDF distribuées entre les sources `human1.rdf` et `human2.rdf`.

5 Conclusion et perspectives

Dans cet article, nous avons présenté une approche pour la publication, le partage et la réutilisation de règles d'inférence sur le Web de données liées. Nous avons choisi des règles SPARQL que nous exploitons dans leur syntaxe SPIN/RDF. Nous avons montré à travers l'utilisation du moteur sémantique Corese/KGRAM que la publication et la réutilisation de règles peuvent ainsi reposer de manière unifiée sur les modèles et techniques du Web de données. La réutilisation de règles repose sur l'interrogation de leur description RDF dans le langage SPARQL. Nous avons présenté différents scénarios de construction automatique de bases de règles pertinentes dans un contexte donné ; nous avons conduit des expérimentations sur la sélection des règles de sémantique de RDFS et de OWL 2 RL pertinentes pour des vocabulaires populaires du Web de données et montré le gain en temps que permet une telle sélection.

Le moteur Corese/KGRAM permet d'implémenter les différents scénarios d'interrogation et d'application de règles multi-sources. Nous avons précisé plusieurs scénarios typiques de réutilisation de règles avec des données liées et nous les avons expérimentés avec un jeu de données de taille réduite. Cependant, l'application de règles d'inférence sur des sources de données distribuées peut s'avérer extrêmement coûteuse. Par exemple, par nature, la base de règles OWL 2 RL est générique. Elle met en jeu certaines règles qui sont très peu sélectives, et qui mènent, dans le cas de sources de données conséquentes à des communications réseau très coûteuses.

Dans la continuité de ces travaux, nous envisageons d'étudier la sélectivité des règles et l'optimisation du moteur de règles (mise en cache) pour proposer des inférence sur des sources multiples de données et de règles dans des temps raisonnables.

Dans la continuation de ces travaux, nous avons commencé à nous intéresser à la traçabilité des règles lorsqu'elles sont sélectionnées sur le Web de données et la mise à jour de bases de règles selon leur contenu ou à les métadonnées qui leur sont associées. Il peut s'agir de modifier

les règles, par exemple en les généralisant en remplaçant une classe ou une propriété par une classe ou une propriété plus générale, de supprimer des règles, d'ajouter des annotations.

Remerciements

Nous remercions Maxime Lefrançois pour l'écriture de la base de règles SPARQL implémentant la sémantique de OWL 2 RL.

Références

ANGLES R. & GUTIERREZ C. (2008). The Expressive Power of SPARQL. In *Proc. of the 7th Int. Semantic Web Conf., ISWC 2008, Karlsruhe, Germany*, volume 5318 of *LNCS*, p. 114–129 : Springer.

BAGET J. F. (2004). Improving the forward chaining algorithm for conceptual graphs rules. In A. PRESS, Ed., *Proc. of the 9th Int. Conf. on the Principles of Knowledge Representation and Reasoning, KR 2004, Whistler, Canada*, p. 407–414.

BIZER C. & SCHULTZ A. (2010). The r2r framework : Publishing and discovering mappings on the web. In *Proc. of the 1st Int. Workshop on Consuming Linked Data, COLD 2010, Shanghai, China* : CEUR-WS.org.

CORBY O., DIENG-KUNTZ R. & FARON-ZUCKER C. (2004). Querying the semantic web with the corese search engine. In *Proc. of the 16th European Conf. on Artificial Intelligence, ECAI 2004, Valencia, Spain*, p. 705–709 : IOS Press.

CORBY O. & FARON-ZUCKER C. (2014). SPARQL Template : un langage de Pretty Printing pour RDF. In *Actes des 25èmes Journées francophones d'Ingénierie des Connaissances, IC 2014, Clermont Ferrand, France*.

CORBY O., GAIGNARD A., FARON-ZUCKER C. & MONTAGNAT J. (2012). KGRAM Versatile Data Graphs Querying and Inference Engine. In *Proc. of IEEE/WIC/ACM Int. Conf. on Web Intelligence, WI 2012, Macau, China* : IEEE Computer Society.

GONZÁLEZ-MORIYÓN G., POLO L., BERRUETA D., TEJO-ALONSO C. & IGLESIAS M. (2012). Assembling rule mashups in the semantic web. In *Proc. of the 9th Extended Semantic Web Conf., ESWC 2012, Heraklion, Crete, Greece*, volume 7295 of *LNCS*, p. 590–602 : Springer.

POLLERES A. (2007). From SPARQL to rules (and back). In *Proceedings of the 16th Int. Conf. on World Wide Web, WWW 2007, Banff, Alberta, Canada*, p. 787–796 : ACM.

POLLERES A., SCHARFFE F. & SCHINDLAUER R. (2007). SPARQL++ for mapping between RDF vocabularies. In *Proc. of the 6th Int. Conf. on Ontologies, DataBases, and Applications of Semantics, ODBASE 2007, Vilamoura, Portugal*, volume 4803 of *LNCS*, p. 878–896 : Springer.

SCHENK S. & STAAB S. (2008). Networked Graphs : a Declarative Mechanism for SPARQL rules, SPARQL Views and RDF Data Integration on the Web. In *Proc. of the 17th Int. Conf. on World Wide Web, WWW 2008, Beijing, China*, p. 585–594 : ACM.

SEYE O., FARON-ZUCKER C., CORBY O. & FOLLENFANT C. (2012). Bridging the Gap between RIF and SPARQL : Implementation of a RIF Dialect with a SPARQL Rule Engine. In *Proc. of Artificial Intelligence meets the Web of Data Workshop at ECAI 2012, Montpellier, France*.

Programmer le web de données avec un « Wiki-based IDE »

Pavel Arapov, Michel Buffa, Amel Ben Othmane

Equipe Wimmics, commune aux laboratoires
INRIA et I3S de Sophia Antipolis,
{arapov, buffa, amel.ben-othmane}@i3s.unice.fr

Résumé : WikiNEXT est un wiki à la croisée des wikis sémantiques et des outils de développement en ligne récents (« web based IDEs »). En permettant de coder directement dans le navigateur des applications exploitant le web de donnée et manipulant des données sémantiques, WikiNEXT étend le concept de wiki sémantique et répond au problème *« comment faciliter la programmation et l'apprentissage d'applications pour le web sémantique ? »*. WikiNEXT s'adresse à plusieurs profils d'utilisateurs, cependant cet article se focalise sur les aspects « wiki programmable » qui concernent principalement les développeurs web. Les wikis sémantiques actuels permettent d'insérer du contenu dynamique dans les pages, mais ne partagent pas cette approche, nous montrerons en quoi WikiNEXT améliore l'état de l'art dans le domaine. L'outil est en ligne sur http://wikinext.gexsoft.com et contient de nombreux tutoriaux interactifs.

Mots-clés : wikis, web-IDE, wikis sémantiques, JavaScript.

1 Introduction

Le problème général auquel WikiNEXT essaie de répondre est *"comment programmer plus facilement des applications exploitant le web de données et les données sémantiques ?"*

Par exemple, nous voulons un graphique de la population des pays n'ayant pas de frontière maritime, et ayant plus de 1.500000 habitants, à partir de données de DBpedia.org (Figure 1).

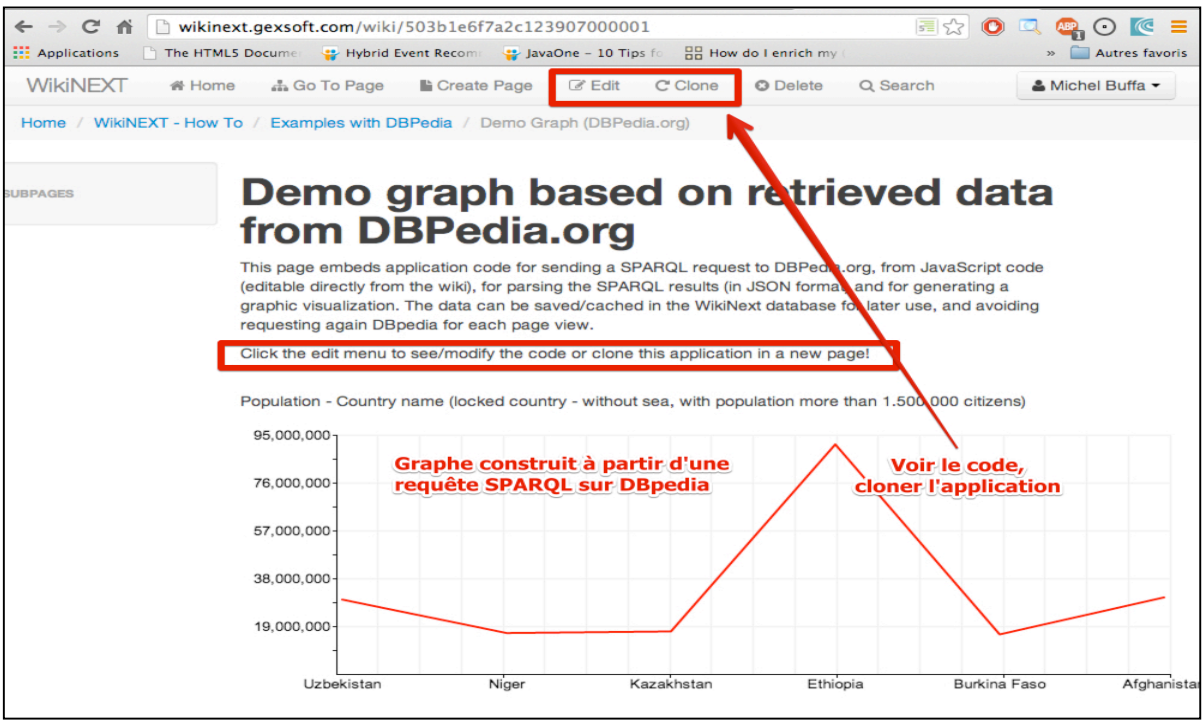
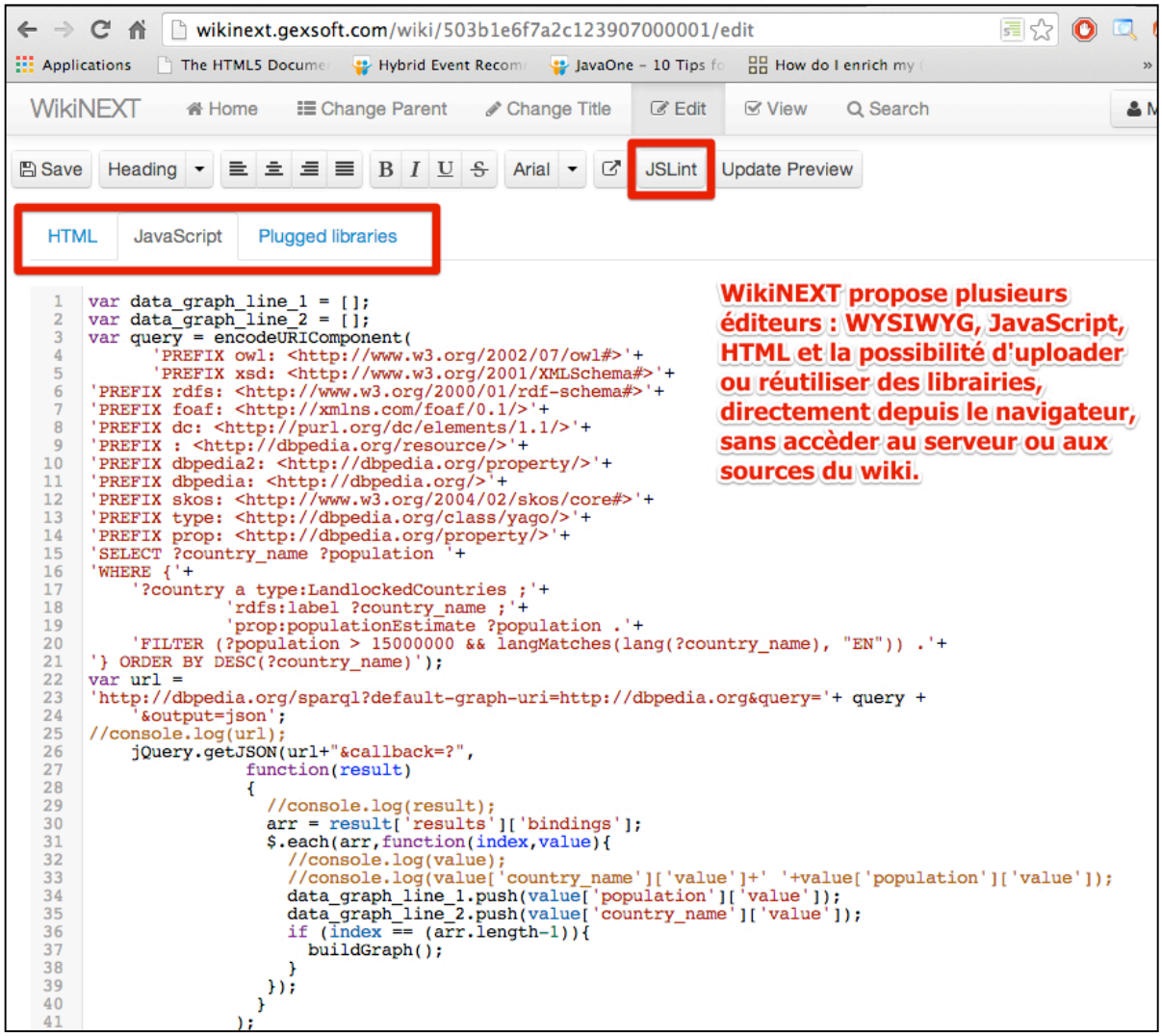


Figure 1 Exemple de contenu dynamique dans WikiNEXT

… on pourra utiliser l'éditeur JavaScript/HTML intégré, qui permet de voir le code de l'application, ou encore, on pourra cloner l'application pour l'examiner et la modifier. Cette approche par mimétisme est caractéristique des wikis et demande peu d'expertise (Figure 2).




```
var data_graph_line_1 = [];
var data_graph_line_2 = [];
var query = encodeURIComponent(
      'PREFIX owl: <http://www.w3.org/2002/07/owl#>'+
      'PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>'+
'PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>'+
'PREFIX foaf: <http://xmlns.com/foaf/0.1/>'+
'PREFIX dc: <http://purl.org/dc/elements/1.1/>'+
'PREFIX : <http://dbpedia.org/resource/>'+
'PREFIX dbpedia2: <http://dbpedia.org/property/>'+
'PREFIX dbpedia: <http://dbpedia.org/>'+
'PREFIX skos: <http://www.w3.org/2004/02/skos/core#>'+
'PREFIX type: <http://dbpedia.org/class/yago/>'+
'PREFIX prop: <http://dbpedia.org/property/>'+
'SELECT ?country_name ?population '+
'WHERE {'+
    '?country a type:LandlockedCountries ;'+
              'rdfs:label ?country_name ;'+
              'prop:populationEstimate ?population .'+
    'FILTER (?population > 15000000 && langMatches(lang(?country_name), "EN")) .'+
'} ORDER BY DESC(?country_name)');
var url =
'http://dbpedia.org/sparql?default-graph-uri=http://dbpedia.org&query='+ query +
    '&output=json';
//console.log(url);
    jQuery.getJSON(url+"&callback=?",
                function(result)
                {
                  //console.log(result);
                  arr = result['results']['bindings'];
                  $.each(arr,function(index,value){
                    //console.log(value);
                    //console.log(value['country_name']['value']+' '+value['population']['value']);
                    data_graph_line_1.push(value['population']['value']);
                    data_graph_line_2.push(value['country_name']['value']);
                    if (index == (arr.length-1)){
                      buildGraph();
                    }
                  });
                }
              );
```

Figure 2 : l'envers du décors de l'application de la Figure 1.

WikiNEXT se programme directement dans le browser et exploite les tendances récentes des outils de développement basés web rendues possible par les avancées de HTML5 et de JavaScript. En proposant des fonctionnalités de clonage et en mélangeant document et application, l'apprentissage par mimétisme se retrouve facilité. WikiNEXT est en développement depuis trois ans (Arapov et Al. 2012), et a été écrit à partir de zéro. Les motivations viennent d'un constat d'expérience : nous avons déjà développé deux wikis sémantiques depuis 2006 : SweetWiki (Buffa et Al. 2008) et DekiWiki (Buffa et Al. 2012), et nous nous sommes heurtés aux limitations que partagent aussi les wikis sémantiques les plus populaires : ils ne permettent pas facilement d'écrire des applications exploitant les données sémantiques et le web de données. La raison principale est le choix d'une approche « par macros » ou « par extensions » au lieu de proposer une interface de programmation dans le navigateur. Pour réaliser l'application présentée, on aurait avec les wikis actuels, inséré une macro dans un document, que l'on aurait pu paramétrer (pour indiquer la requête, le type de graphe, etc.), mais le code « derrière » une macro nécessite des outils *ad hoc* lourds et un accès au serveur. Seuls des experts peuvent s'attaquer à ce type de développement.

L'originalité de WikiNEXT est qu'il est à la fois un wiki sémantique, et un IDE[1] (Integrated Development Environment), il considère les pages comme des applications contenant du texte, des images ou autres médias, des métadonnées et aussi une partie programmable. Les avantages par rapport à l'approche basée « macros » sont nombreux :

- *Programmation instantanée* (« live coding ») : je code et je vois le résultat quasiment en temps réel de mes modifications, cette approche facilite les expérimentations. Elle est très adaptée à des applications pédagogiques,
- *Partage des applications créées dans le wiki* (clonage, partage),
- *Documentation « sur place » des applications* : une page du wiki peut contenir un tutorial ou la documentation de sa partie applicative,
- *Utilisation de langages standards (JavaScript, JSON, RDF, RDFa, SPARQL, etc.).*

WikiNEXT propose également des outils pour aider le développeur : il fournit une riche API accessible depuis le code JavaScript, des aides interactives (pour la création de requêtes SPARQL par exemple), un système de templates pour l'affichage de données sémantiques, une base de données de graphes compatible SPARQL 1.1 pour le raisonnement et pour l'implémentation d'un SPARQL endpoint, une base de données objet pour la persistance, etc.

L'outil est en ligne[2], open source, et propose de nombreux exemples et tutoriaux montrant comment collecter des données depuis DBPedia, depuis Freebase, ou depuis la base de connaissance locale du Wiki. Les démonstrations et tutoriaux sont variés : mashup de plusieurs sources de données, annotation automatique de documents, génération de vues personnalisées sous forme de page de wiki ou sous forme de visualisations graphiques (templates), utilisation d'un cache permettant de ne pas requêter le web de données à chaque affichage de page enrichie, avec gestion de la durée de vie des données, etc.

WikiNEXT s'adresse à plusieurs profils d'utilisateurs. Un utilisateur lambda peut utiliser l'éditeur WYSIWYG et créer des documents classiques, cependant nous avons mis dans cet article l'accent sur le côté « wiki pour écrire des applications ». Nous nous adressons donc aux développeurs d'application web et web sémantique, mais aussi aux enseignants et à ceux qui utilisent déjà des petits « web IDEs » pour tester du code.

Etat de l'art : convergence entre wikis classiques, wikis d'application, wikis sémantiques et « environnement de développement basés web »

Wikis d'application : Issus du Web 2.0, les wikis, dont le représentant le plus illustre est WikiPedia, sont connus pour favoriser la convergence de l'écriture collaborative de documents. Rapidement, certains moteurs de wikis ont étendu cette approche pour aller au-delà de l'écriture de simples documents textuels. Les « wikis d'application », aussi appelés « wikis d'entreprise », ont ajouté le moyen de « programmer dans les pages du wiki ». TWiki (encore populaire de nos jours), en 1998, a été un pionnier de cette approche en proposant d'utiliser des macros pour générer du contenu dynamique dans les pages. Par exemple on pouvait générer des tableaux, des formulaires, des champs des recherche dans les pages, les données manipulées étant elles aussi stockées dans des pages ayant une structure particulière (sous forme de tables ou de liste HTML). En 2005, d'autres wikis, comme JOT (devenu par la suite Google Sites), XWiki ou Mindtouch Core[3] sont allés plus loin en intégrant de véritables langages de script. Ainsi des utilisateurs avancés ayant des notions de programmation pouvaient développer des « petites applications dans le wiki » incluant la définition de modèles de données, leur traitement et leur présentation. Tout en conservant la possibilité pour d'autres utilisateurs de copier/coller une page/application et de modifier légèrement le travail cloné pour en faire une version personnelle. C'est ce que Ward Cunningham appelle *« the Wiki Way »* et qui transforme les wikis en puissants outils pour *« apprendre en*

[1] IDE = Integrated Development Environment. Eclipse ou Netbeans sont des IDEs classiques, jsbin.com, jsfiddle.net sont des IDEs en ligne, ou « web based IDEs ».

[2] http://wikinext.gexsoft.com.

[3] http://www.mindtouch.com/, utilise DekisScript, un langage inspiré par JavaScript, pour écrire du code dans les pages.

regardant le travail des autres » (Leuf et Cunningham 2001). Le code source de ces applications et les données ne sont pas visibles par les utilisateurs qui consultent les pages, mais éditer une page révèle les secrets de fabrication (le script, les macros) de l'application qu'elle renferme. Ces macros sont soit prédéfinies, soit écrites par les utilisateurs à l'aide d'un langage de script (Velocity[4] ou Groovy[5] pour XWiki, DekiScript, pour le wiki Mindtouch Core, etc.) Les ensembles de macros peuvent utiliser des plugins ou des extensions ajoutés côté serveur (par exemple un plugin pour se connecter à une base de données SQL externe, fournira une macro utilisable dans les pages). *Dans ce cas, le développeur du plugin ou de l'extension doit avoir accès au code source du Wiki, ou à un SDK fourni avec le moteur de wiki. Parfois le nombre d'outils nécessaires pour pouvoir développer un plugin est très important.*

Les Wikis sémantiques : inspirés par les wikis traditionnels, les chercheurs ont commencé à partir de 2005 à proposer des « wikis sémantiques » tels que Semantic Media Wiki (Krötzsch et Vrandecic 2006), IkeWIki (Shaffert 2006) ou encore SweetWiki (Buffa et al. 2008), voir (Buffa et al. 2008) et (Krotzsh et al. 2007) pour une synthèse des moteurs de wiki sémantiques de la première génération (2005-2008). Ces wikis permettent à leurs utilisateurs d'ajouter du contexte sémantique aux documents tout en préservant la simplicité et l'essence collaborative des wikis. Les ressources sémantiques sont en général conservées en interne (dans une base de connaissance) sous la forme de données RDF, OWL ou sous forme de graphes conceptuels (Oren, Breslin, Decker 2006). Les données sémantiques produites sont déjà intéressantes en elles-mêmes (par exemple dans le cas de production collaborative d'ontologie, ou d'annotation de documents), dans la mesure où elles sont naturellement difficiles à produire, mais surtout, les wikis équipés de moteurs sémantiques peuvent les exploiter pour du raisonnement, implémenter des outils de recherche augmentée, des systèmes de suggestion, etc. Les wikis sémantiques proposent souvent des langages intermédiaires proches du langage naturel (on écrit « this is a class » dans Semantic Media Wiki pour créer une classe qui sera un concept dans une ontologie) ou à base de langage de markup popularisé par la « culture wiki ». Semantic Media Wiki, ou SemperWiki (Oren 2005) utilisent une syntaxe à base de crochets pour intégrer des annotations, qui seront ensuite traduites en RDF une fois la page sauvegardée. Cette approche, très simple, basée sur les usages des wikis de l'époque, a posé quelques problèmes de cohérence dans les premières générations de wikis sémantiques (2005-2010). IkeWIki (Shaffert 2006) a combiné l'approche classique des wikis pour la création de documents avec une interface à base de formulaires avec auto-complétion pour faciliter la réutilisation de vocabulaires et la vérification de cohérence. OntoWiki (Auer et al. 2006) quant à lui est un véritable éditeur d'ontologie avec une interface utilisateur proposant différentes vues pour naviguer et organiser l'ontologie produite. SweetWiki (Buffa et al. 2008) proposait un système de tagging social sémantique et conceptualisait le wiki lui-même sous la forme d'une « ontologie du wiki », implémentant ainsi la quasi-totalité de ses fonctionnalités autours des technologies du web sémantique.

Aujourd'hui, peu de wikis sémantiques sont encore maintenus[6] on pourra retrouver un état de l'art récent dans (Meilendera et al. 2010). Existent encore des wikis de l'ancienne génération qui ont (1) étendu leurs fonctionnalités en devenant des hybrides de CMS (Content Management Systems) ou (2) au contraire en se « coupant en morceaux » et en externalisant certaines fonctionnalités à des entités externes dédiées. Dans la première catégorie nous retrouvons Semantic Media Wiki avec aujourd'hui de très nombreuses extensions, comme Halo, qui propose des formulaires d'annotation avec auto-complétion, un éditeur WYSIWYG, l'intégration de fichiers multimédia et la mise en place d'un « SPARQL endpoint ». On trouve aussi des wikis basés sur Semantic Media Wiki, comme Moki (Rospocher et al 2009),

[4] http://velocity.apache.org/

[5] http://groovy.codehaus.org/

[6] Il y en avait 37 de recensés lors du workshop qui leur était dédié lors de la conférence ESCW 2006, moins d'une dizaine sont encore actifs en 2013.

spécialisé sur la modélisation de processus d'entreprises, ou OWiki (Di Lorio et al. 2010), dédié à la génération de contenu dirigé par des ontologies. Dans la seconde catégorie on trouve le projet KiWi -Knowledge In Wiki- (Shaffert et al. 2008), qui implique les auteurs de IkeWiki, refondu dans un projet comprenant de nombreux modules, ou SweetDeki (Buffa et al. 2013, Buffa et Husson 2012), le wiki qui a succédé à SweetWiki, fondu dans le projet ANR ISICIL (Gandon et al. 2009) intégrant un réseau social, un serveur de tags, un gestionnaire de ressources sémantiques externes, etc. Ces projets ont essayé de mettre le wiki sémantique dans un contexte industriel (ISICIL a donné naissance à la société Mnemotix.com, KiWi est appuyé par des fabricants de CMSs). Bien que les wikis sémantiques cités permettent la création d'ontologie ou l'annotation de documents, l'exploitation des données sémantiques par l'utilisateur se limite souvent à l'insertion de requêtes ou macros dans les pages, faisant références à des extensions ou plugins éventuellement installés. *Dans ce sens, ces wikis sémantiques modernes ont convergé vers les wikis d'applications « historiques » présentés dans la précédente section, et partagent également leurs défauts : développer un nouveau plugin ou une extension requiert un SDK ou l'accès aux sources du wiki. On peut développer de petites applications « dans le wiki », mais avec des contraintes importantes.*

Outils de développement basés web (*« Web-based IDEs »*) : L'idée consistant à écrire des applications directement dans un navigateur web n'est pas récente : les premières expérimentations datent de 1996 (Crespo et al. 1996). Depuis 2005, avec la technologie Ajax et le développement de JavaScript, les éditeurs de code source pouvant fonctionner dans une page web n'ont cessé de s'améliorer. Des éditeurs comme Code Mirror[7], ACE Cloud9[8] or ternjs[9] proposent aujourd'hui la colorisation de syntaxe, l'auto-complétion, des fonctions d'édition complètes, et équipent plusieurs sites web connus proposant des environnements de développement basés web[10]. Certains supportent plusieurs langages de programmation comme compile-online.com, ideone.com ou compilr.com, mais les plus intéressants sont sans doute les environnements dédiés à la programmation JavaScript/HTML/CSS, très appréciés des *développeurs web*. Des sites comme jsbin.com, jsfiddle.net, codepen.io ou tinkerbin.com proposent d'écrire du code directement dans le navigateur et de voir l'exécution en temps réel, pendant la saisie. On parle de « live coding », ou « codage vivant », qui est particulièrement appréciable pour évaluer rapidement des algorithmes, des idées, ou tout simplement pour tester des fonctions d'APIs de HTML5. Ces outils proposent souvent une option de clonage d'une application existante. La plupart de ces outils, cependant, ne sont que des bacs à sable pour tester du code ou pour écrire des exemples à but pédagogique. Enfin, on trouve des outils de développement comme cloud9ide.com, shiftedit.net, coderun.com ou codeanywhere.net qui ambitionnent de remplacer des outils comme Eclipse, en proposant de gérer en ligne le développement de véritables projets composés de multiples fichiers JavaScript/HTML/CSS. WikiNEXT reprends certaines fonctionnalités de la dernière génération des outils présentés dans cette section. A notre connaissance, aucun wiki sémantique n'a encore adopté une telle approche pour « programmer dans le wiki ».

2 WikiNEXT

Motivation : wikis d'application, wikis sémantiques, IDEs basés web : WikiNEXT est un hybride, un mélange de ces trois approches. Tout d'abord, dans WikiNEXT, *chaque page est une application web versionnée*, composée de code JavaScript/HTML/CSS. Nous l'avons vu dans l'introduction, plusieurs éditeurs sont fournis : un éditeur WYSIWYG pour des documents classiques, et un éditeur de code pour la partie « application ». Par ailleurs WikiNEXT expose une API donnant aux applications accès aux mécanismes internes du wiki,

[7] http://codemirror.net/

[8] http://ace.c9.io/

[9] http://ternjs.net/

[10] See http://en.wikipedia.org/wiki/Comparison_of_JavaScript-based_source_code_editors

et facilitant la manipulation de données sémantiques (requêtes, affichage, persistance). Considérons par exemple une page classique créée par un utilisateur ne sachant pas programmer. Il pourra demander à une personne ayant des compétences en développement web « d'augmenter sa page » en écrivant un bout de code qui collecte des informations d'une source de données externe comme DBPedia afin d'annoter le document qu'il a écrit, ou pour extraire des éléments du texte afin de suggérer une classification, insérer une visualisation graphique, ou sauvegarder les données collectées dans la base de connaissances du wiki. Il sera possible par la suite de cloner la page/application existante pour l'étudier et en faire une version différente. Comme dans tout wiki, il est nécessaire d'amorcer le wiki en proposant une base de données initiale d'exemples variés, de tutoriaux, afin de faciliter l'apprentissage par clonage/modification. WikiNEXT propose des services pour l'annotation de textes, la persistance « back end », côté serveur, en fournissant notamment une base de données de graphe compatible RDF/SPARQL 1.1, et une base de données objet, WikiNEXT est aussi un SPARQL endpoint. Nous avons mis l'accent sur cet aspect « programmation » du wiki, car les applications sémantiques sont souvent associées avec des « documents » qui sont soit sources de données, soit cibles pour des annotations ou des enrichissements. Par ailleurs, WikiNEXT a aussi une vocation pédagogique, en tant qu'outil permettant de développer des tutoriaux interactifs. On peut, par exemple, écrire un document avec du texte explicatif, mélangé avec l'application qu'il décrit : un tutorial sur « comment requêter DBPedia » contiendra à la fois le texte et le code de l'application, présentera les résultats d'exécution en temps réel et proposera de modifier directement le code pour voir le résultat des modifications.

La couche « application » d'un document est constituée de code JavaScript qui peut être édité de la même manière que l'on peut éditer le code HTML de la page. Ainsi il est possible de manipuler depuis le code JavaScript le Document Object Model (DOM) de la page comme on le fait dans la programmation JavaScript classique : rajouter un id sur un HTML pour afficher des résultats, ou bien utiliser le système de templates de présentation de WikiNEXT, présenté dans la section suivante. Certaines tâches comme « créer une nouvelle page du wiki », « requêter la base de connaissance du wiki », « annoter un document et maintenir les vues sur ces annotations à jour lorsque les valeurs changent » (liste non exhaustive) sont réalisées par des fonctions de l'API interne du wiki, alors que d'autres peuvent s'appuyer sur des librairies JavaScript externes telles que D3.js[11], une puissante librairie graphique pour la visualisation de données. WikiNEXT propose par ailleurs un outil intégré permettant d'uploader des librairies externes et de sélectionner celles utilisées par chaque page/application. Le prototype est open source[12] et disponible en ligne[13]. Il inclut aujourd'hui de nombreux exemples et tutoriaux. Nous l'utilisons avec des étudiants de Master pour illustrer certaines utilisations du web de données et pour apprendre les langages RDF/SPARQL.

Architecture du logiciel : dans le paradigme de développement agile, des chercheurs ont montré que les wikis peuvent être utilisés comme plate-forme pour l'échange d'objets logiciels légers réutilisables (Rech, Bogner et al. 2007), cependant il s'agissait de « bouts de logiciels » externes au wiki. WikiNEXT reprend ces principes et les mets en œuvre dans le paradigme MVC, la partie JavaScript de la page formant « la couche métier et le contrôleur », la partie « vue » se faisant dans le code HTML / CSS de la page, et la partie « modèle » étant composée de métadonnées RDF. La synchronisation de la vue et des modèles est implémentée par WikiNEXT. Il est également possible de synchroniser les modèles avec les bases de données de WikiNEXT, situées côté serveur, via l'API WikiNEXT. Chaque page WikiNEXT est associée avec un ensemble de métadonnées qui décrivent ses principales caractéristiques : titre, auteur, contributeurs, dernière date de modification, version, etc. Mais les pages sont

[11] http://d3js.org

[12] https://github.com/pavel-arapov/wikinext

[13] http://wikinext.gexsoft.com

aussi des « containers » : elle contiennent des métadonnées qui auront pu être rajoutées manuellement ou par une application d'annotation. WikiNEXT s'appuie sur les ontologies de schema.org[14] pour décrire la structure des pages, les utilisateurs, etc. Chaque page est l'instance d'un Article, elle est représentée par un graphe nommé constitué d'annotations RDFa/RDFAlite[15], dont le nom est basé sur l'URI de la page (Figure 3). Nous appelons l'ensemble des métadonnées localisées dans une page la *base de connaissances locale*. L'ensemble des pages du wiki, est appelé *base de connaissances globale*, elle est constituée de l'union de tous les graphes nommés. Les métadonnées générées par la partie programmable d'une page, par exemple une liste de pays provenant d'une requête sur DBPEdia.org, pourront être injectées dans la page sous forme d'annotations RDFa/RDFalite. Elles seront également sauvegardées dans un triple store intégré à la partie serveur du Wiki lors de la sauvegarde de la page.

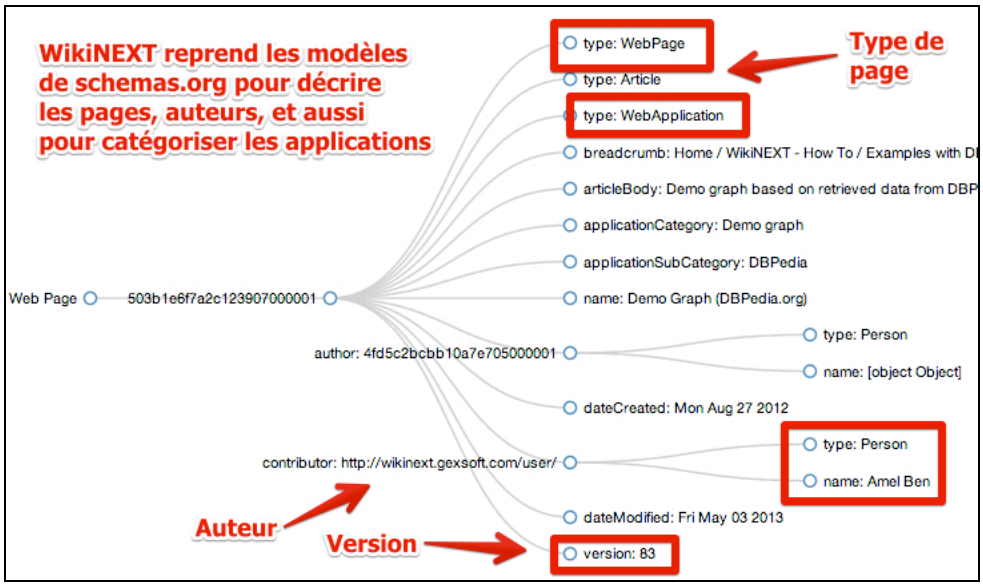


Figure 3: Le modèle des pages de WikiNEXT

Implémentation (Figure 4**) :** WikiNEXT a été écrit « from scratch » en JavaScript. Il s'appuie sur le micro-serveur Node.js[16] et sur la base de données MongoDB[17]. Node.js embarque l'interpréteur JavaScript V8 de Google, qui équipe entre autre le navigateur Google Chrome. Ainsi, Node.js permet ce que l'on appelle le « server-side JavaScript », une tendance émergente dans le développement web. MongoDB est une base de données NoSQL orientée documents (indexation du contenu, etc.). Au lieu de stocker des objets sérialisés sous forme de lignes et colonnes dans des tables comme on le ferait avec une base de données relationnelle traditionnelle, ici on stocke des objets JavaScript au format BSON (une forme binaire du format de sérialisation JSON[18] de JavaScript). Voir Figure 4. Nous utilisons pour gérer les triplets RDF une version améliorée de RDFStore-js (Hernandez et Garcia 2012), un moteur de graphes compatible SPARQL 1.1, écrit en JavaScript. Le contenu traditionnel de la page est stocké sous forme d'objets dans la base de données MongoDB (notamment, nous générons des indexes pour le contenu textuel de la page). Nous l'utilisons également comme couche de persistance pour RDFStore-js. Ainsi, les triplets sont accessibles à la fois à travers

[14] http://schema.org/ and http://schema-rdfs.org ont été créés par les principaux moteurs de recherche (Google, Microsoft, Yahoo et Yandex). Ils prooposent des vocabulaires partageables RDF/S et pour les microdatas de HTML5, qui couvrent les principaux domaines.

[15] http://www.w3.org/TR/xhtml-rdfa-primer/ et http://www.w3.org/TR/rdfa-lite/

[16] http://nodejs.org

[17] http://mongodb.org

[18] http://www.json.org, a text based notation for JavaScript objects.

des requêtes SPARQL, mais aussi directement via MongoDB. On dispose d'un accès direct aux nœuds du graphe, ce qui est utile pour implémenter de manière très efficace certaines tâches (recherche d'entités nommées, notamment). Pour l'affichage et la visualisation d'annotations sémantiques, WikiNEXT propose un système de templates. Le cas d'utilisation classique consiste à créer des pages WikiNEXT faisant office de modèles de présentation (templates) qui seront réutilisées par d'autres pages, par leur partie applicative. Le principe est simple et permet de découpler l'interface utilisateur de la définition des données.

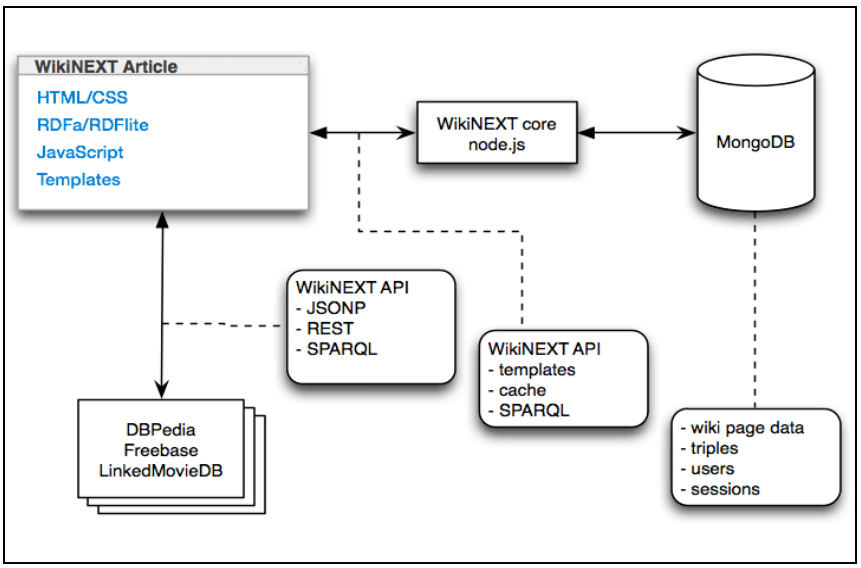


Figure 4: Architecture technologique de WikiNEXT

Les templates, la manipulation de requêtes SPARQL, la manipulation d'annotations RDF, la persistance, le cache des données, toutes ces fonctionnalités sont accessibles par le code des applications que l'on écrit dans le wiki, au travers de l'API WikiNEXT. Avec ces choix technologiques, nous avons essayé de minimiser le nombre de langages de programmation utilisés et le nombre de formats de données. Avec WikiNEXT, JavaScript est utilisé le long de la chaîne complète, depuis le code fonctionnant dans le navigateur web (code client) jusqu'au code serveur et au format de persistance.

3 Mashup sémantique et utilisation de la base de connaissance globale de WikiNEXT

WikiNEXT permet de développer des applications plus complètes et plus variées que l'exemple très simple présenté dans l'introduction. Nous décrivons ici un scénario d'usage complet : le développement d'une application qui utilise le SPARQL endpoint de DBPedia.org pour récupérer des données RDF concernant des villes (description, population, photos, etc.). Les résultats sont utilisés pour créer à la volée des pages du wiki –une par ville-, basée sur un modèle de présentation, lui aussi créé dans le wiki. Les pages sont annotées et les annotations sont sauvegardées dans le triple store RDF du wiki. Le nom des villes à chercher se fait à l'aide d'un formulaire, lui aussi créé dans le wiki ; enfin, une dernière page réalise un « mashup sémantique » en construisant une carte présentant l'ensemble des villes qui ont été récupérées, avec un résumé et des photos de chacune d'entre elles. Cette application est disponible en ligne sur le site de WikiNEXT[19], c'est un des tutoriaux que l'on peut consulter, éditer et dont on peut voir le code, le modifier ou le cloner[20]. Elle est également disponible en

[19] http://wikinext.gexsoft.com/wiki/519e04c580194c4178000001

[20] Il suffit de se connecter au Wiki (inscription traditionnelle ou via identifiants facebook) et d'éditer la page de l'application « City ».

screencast[21] sur YouTube. Le tutorial guide pas à pas le développeur : on commence par créer en HTML un formulaire de saisie (Figure 5) permettant de choisir la ville à rechercher, le pays, la langue. Les paramètres saisis servent à requêter le SPARQL endpoint de DBPEdia.org. Les résultats sont ensuite utilisés pour créer à la volée une page dans le wiki. Pour cette application on ne va pas conserver le modèle original de DBPedia.org pour les villes mais on va décrire les villes en utilisant des propriétés du modèle http://schema.org/city, plus simple (Figure 7). Lorsqu'une page du wiki est sauvegardée (c'est le cas de la page créée à la volée pour chaque ville-résultat, cf Figure 6), les métadonnées sont également ajoutée à la base de connaissance de WikiNEXT (au format RDF). L'intérêt de cet exemple est de montrer l'utilisation des fonctions d'API pour manipuler les annotations sémantiques et pour illustrer l'utilisation de la base de connaissance RDF/SPARQL de WikiNEXT.

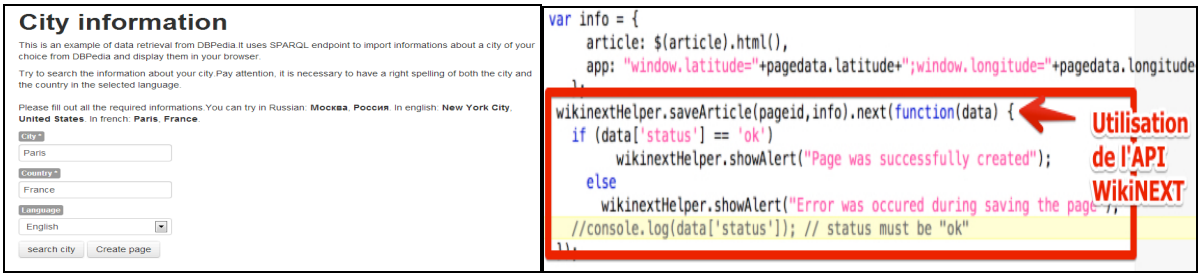


Figure 5: Formulaire de saisie de l'application « City », un des tutoriaux de WikiNEXT, et exemple de code utilisant l'API de WikiNEXT pour créer une page à la volée avec les résultats d'une requête SPARQL.

L'API fournit d'autres fonctionnalités comme la possibilité de cacher/mémoriser des données avec une certaine durée de vie. Par exemple, la page de la ville de Paris contient des annotations concernant la population, mais on peut indiquer que ces données ont une certaine durée de vie. Passé un délai, l'affichage de la page déclenchera un rafraichissement des données depuis DBPedia.org (au lieu de les prendre dans le cache local, c'est-à-dire dans la base de connaissance locale, voir un des exemples en ligne disponibles[22]).

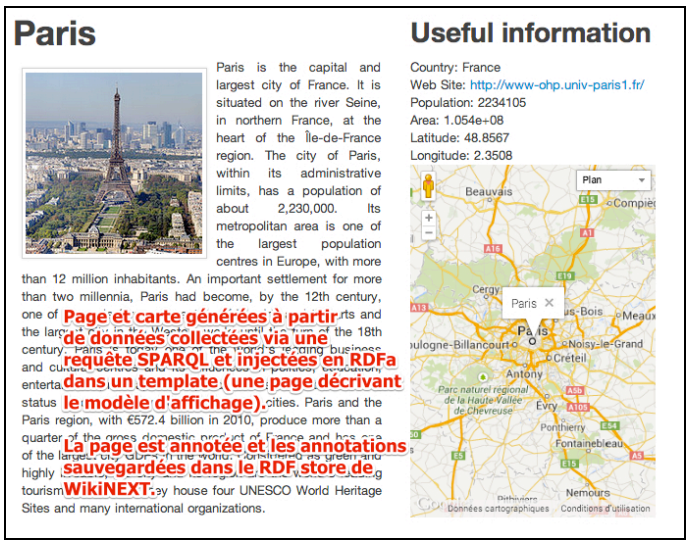


Figure 6: Exemple de page générée

[21] Chercher « WikiNEXT » sur Youtube.com

[22] Exemple d'application utilisant un cache avec durée de vie : http://wikinext.gexsoft.com/wiki/511a4f8207d1b7e451000001

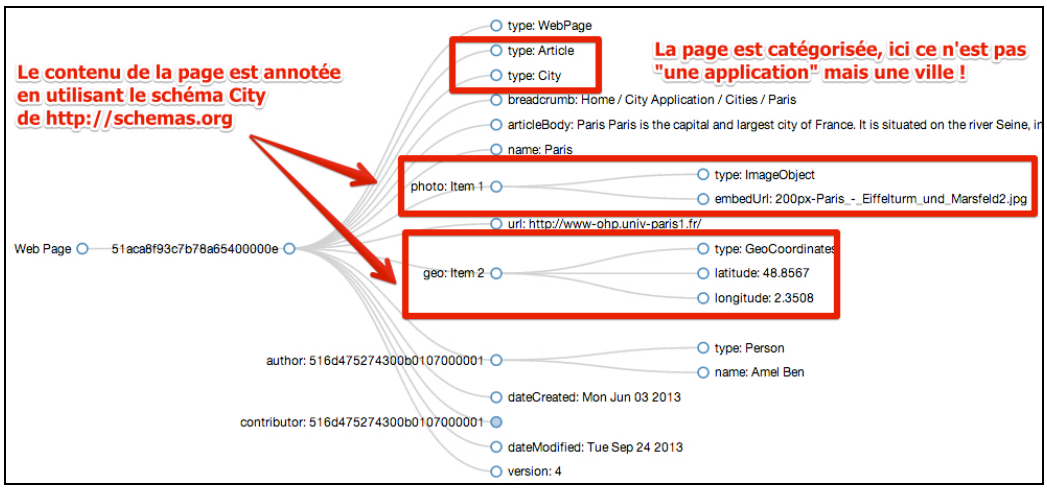
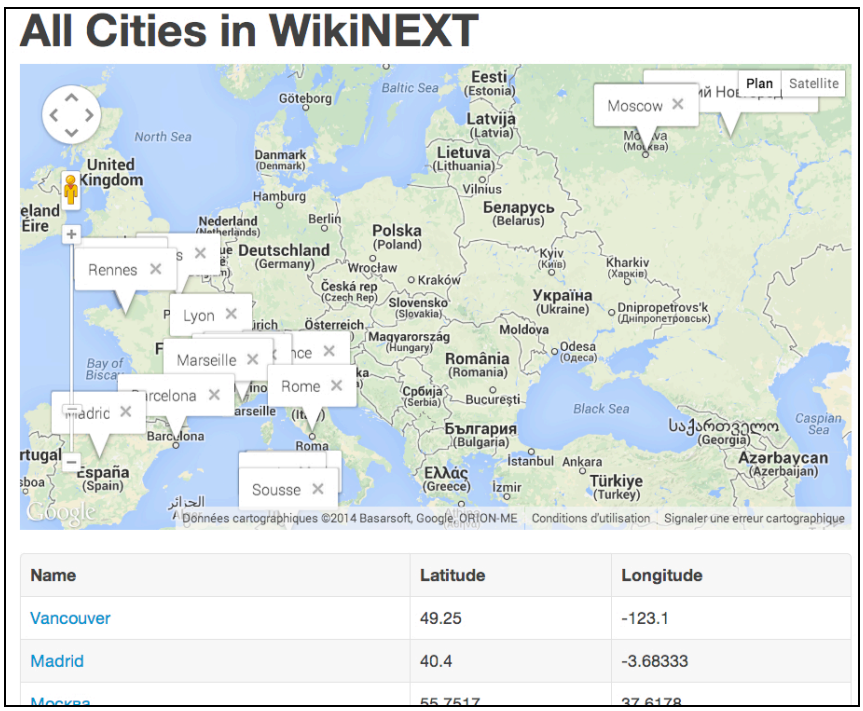


Figure 7: Annotations d'une page créée dynamiquement, pour la ville de Paris.

Réutilisation des données : pour illustrer l'utilisation de la base de connaissance globale de WikiNEXT, nous pouvons également réaliser un mashup à partir des métadonnées associées aux pages des villes construites à partir de requêtes envoyées depuis le formulaire présenté précédemment. Rappelons que les métadonnées correspondant à chaque ville sont à la fois présentes dans les pages mais également dans le triple store de WikiNEXT. Pour l'application de mashup (une carte avec l'ensemble des villes et une liste contenant une vue résumée de chaque ville), nous créons une nouvelle application/page dans le wiki mais cette fois-ci nous requêtons en SPARQL la base de connaissance du wiki au lieu de requêter le web de données, les résultats sont présentés Figure 8. WikiNEXT pré-charge les ontologies de schema.org, les données sur les villes stockées dans WikiNEXT ont été, dans cet exemple, écrites en se référant au schéma http://schemas.org/City.



Name	Latitude	Longitude
Vancouver	49.25	-123.1
Madrid	40.4	-3.68333
Москва	55.7517	37.6178

Figure 8 : Mashup sémantique à partir de la base de connaissance de WikiNEXT

4 Usages et évaluation de WikiNEXT

En développement depuis trois ans, WikiNEXT a été développé principalement par une personne dans le cadre de sa thèse, mais la nature ouverte du wiki fait qu'il a été testé par un groupe de 5 à 6 personnes au cours de son développement. La plupart des exemples en ligne ont été développés par ce noyau de béta testeurs. Il y a quelques mois l'outil est devenu stable et nous avons procédé à des évaluations auprès de personnes n'ayant pas participé à son développement. Nous avons défini un protocole de test consistant à effectuer divers tutoriaux, notamment l'application décrite dans la section 4. Le groupe de test était constitué de développeurs web et d'étudiants en master ou en thèse d'informatique. Nous avons mesuré le temps passé pour chaque étape, observé les comportements et noté tous les problèmes rencontrés. Finalement nous avons conduit des interviews avec chaque personne et demandé de comparer cette expérience avec un développement classique pour effectuer les mêmes tâches. Nos web développeurs ont suivi un cours sur le web sémantique et ont une petite expérience de développement dans ce domaine. Les résultats[23] ont montré que WikiNEXT est jugé simple à utiliser et propose une manière originale et rapide pour requêter le web de données et réutiliser les données sémantiques. La principale difficulté rencontrée par les utilisateurs a été l'écriture de requêtes SPARQL : la syntaxe en elle-même a parfois posé problème, notamment la nécessité d'encadrer certains caractères avec des anti slash en JavaScript, mais aussi l'apprentissage de vocabulaires nouveaux. Nous avons depuis écrit des outils d'aide au requêtage (traitement automatique des caractères spéciaux pour JavaScript, widget avec des requêtes prêtes à l'emploi, historique des requêtes récentes, auto-complétion sur les propriétés des vocabulaires, etc.). Ces applications[24] d'aide ont toutes été écrites dans WikiNEXT, ce qui montre son potentiel.

Des expérimentations à plus grande échelle, nécessaires, sont en cours auprès d'étudiants qui suivent le cours Web Sémantique en master 2 informatiques à l'Université de Nice, et avec les étudiants internationaux qui suivent le cours HTML5 du W3C[25]. Par ailleurs nous développons actuellement une application de synthèse des activités des équipes de l'INRIA Sophia-Antipolis en exploitant le rapport d'activité de l'année 2012 disponible dans une version RDF/JSON. Une telle application a déjà été développée à l'aide de Semantic Media Wiki, ce qui constituera une bonne base de comparaison. Les auteurs de cette application sont impliqués dans la nouvelle expérimentation.

5 Discussion et conclusion

WikiNEXT propose une approche originale en mélangeant wiki sémantique et environnement de développement basé web (« web-based IDE »). S'appuyant sur des technologies émergentes, il propose aujourd'hui un environnement intégré pour programmer le web de données. Ses applications sont principalement pédagogiques (pour enseigner le web de données ou HTML5) mais aussi, il concurrence directement les outils de développement basés web comme jsbin.com ou jsfiddle.net, en ajoutant un moteur de templates, une base de connaissances, une puissante API pour manipuler des données sémantiques et pour requêter le web de données. Des tests d'usage ont été menés auprès de développeurs web et d'étudiants de master informatique, avec succès. En ajoutant l'aspect « programmation d'application dans la navigateur » il étend le concept de wiki sémantique en ouvrant de nouveaux champs d'application. L'outil est en ligne et propose plusieurs tutoriaux. Nous poursuivons son développement et allons procéder cette année à de nouvelles évaluations.

[23] Le protocole de test, les résultats, le questionnaire, sont disponibles : http://wikinext.gexsoft.com/wiki/52304d287a61be9c29000011

[24] Voir par exemple l'outil d'aide au requêtage SPARQL : http://wikinext.gexsoft.com/wiki/51d540a7528188fb3c000038

[25] Le cours HTML5 du W3C est disponible sur http://w3devcampus.com. Il a été écrit par un des auteurs de cet article. WikiNEXT a été utilisé pour écrire certains exemples interactifs et tutoriaux.

Références

AUER, S., DIETZOLD, S., & RIECHERT, T. (2006). OntoWiki–A tool for social, semantic collaboration. In The Semantic Web-ISWC 2006 (pp. 736-749). Springer Berlin Heidelberg.

ARAPOV, P., & BUFFA, M. (2012, April). WikiNext, a JavaScript semantic wiki. In developer track, WWW2012 Conference, Lyon.

ARAPOV, P., & BUFFA, M. (2012, August). WikiNext, a JavaScript wiki with semantic features. In Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration (p. 39). ACM.

BUFFA, M. (2005) Michel Buffa "Intranet Wikis", workshop "intrawebs" of WWW 2006, Edinburgh, Scotland.

BUFFA, GANDON, ERETEO, SANDER, & FARON (2008), SweetWiki: A semantic wiki, Special Issue of the Journal of Web Semantics on Semantic Web and Web 2.0, Volume 6, Issue 1, February 2008 , Edited by Mark Greaves and Peter Mika, Elsevier, Pages 84-97

BUFFA, M., HUSSON, G. (2012) : « SweetDeki : le wiki sémantique couteau suisse du réseau social ISICIL », 12ème Conférence Internationale Francophone "Extraction et Gestion des Connaissances" EGC'12, Janvier 2012, Bordeaux, pages 387-399

CRESPO, ARTURO & BIER, ERIC A. (1996) WebWriter: A browser-based editor for constructing web applications. Computer Networks and ISDN Systems, 1996, vol. 28, no 7, p. 1291-1306.

DI IORIO, A., MUSETTI, A., PERONI, S., VITALI, F. (2010) : Ontology-driven generation of wiki content and interfaces. The New Review of Hypermedia and Multimedia 16(1&2): 9-31 (2010)

OREN E. SemperWiki: a semantic personal Wiki. In SemDesk. 2005

GANDON, ABDESSALEM, BUFFA, & AL. (2009), ISICIL: Information Semantic Integration through Communities of Intelligence online, 10th IFIP Working Conference on Virtual Enterprises, Thessaloniki, Greece, 7-9 October 2009.

KRÖTZSCH. M., VRANDEČIĆ. D.. & VÖLKEL. M. (2006). Semantic mediawiki. In The Semantic Web-ISWC 2006 (pp. 935-942). Springer Berlin Heidelberg.

KROTZSCH, MARKUS, SCHAFFERT, SEBASTIAN, & VRANDECIC (2007), DENNY. Reasoning in semantic wikis. In: Reasoning Web. Springer Berlin Heidelberg, 2007. p. 310-329.

HERNANDEZ A. G. AND GARCIA M. N. M. A JavaScript RDF store and application library for linked data client applications. In Devtracks of the, WWW2012, conference. Lyon, France. 2012

LEUF, B., & CUNNINGHAM, W. (2001). The Wiki way: quick collaboration on the Web, Book.

MEILENDERA, T., JAYB, N., LIEBERB, J., & PALOMARESA, F (2010). Semantic wiki engines: a state of the art. Semantic-web-journal.net. IOS Press, 2010.

OREN, E., BRESLIN, J. G., & DECKER, S. (2006, May). How semantics make better wikis. In Proceedings of the 15th international conference on World Wide Web (pp. 1071-1072). ACM

RECH, J., BOGNER, C., & HAAS, V. (2007). Using wikis to tackle reuse in software projects. Software, IEEE, 24(6), 99-104.

ROSPOCHER, M., GHIDINI, C., PAMMER, V., SERAFINI, L., LINDSTAEDT, S.: (2009) MoKi: the Modelling wiKi. SemWiki 2009

SCHAFFERT, S. (2006, June). IkeWiki: A semantic wiki for collaborative knowledge management. In Enabling Technologies: Infrastructure for Collaborative Enterprises, 2006. WETICE'06. 15th IEEE International Workshops on (pp. 388-396). IEEE.

SCHAFFERT, S., EDER, J., SAMWALD, M. & BLUMAUER, A. (2008). Kiwi - knowledge in a wiki. In European Semantic Web Conference 2008.

Posters et démonstrations

Une Plateforme Support à l'Apprentissage Organisationnel

Ala Atrash, Marie-hélène Abel, Claude Moulin

HEUDIASYC, UMR CNRS 7253, Université de Technologie de Compiègne, BP 20529, 60205, Compiègne, France
{ala-aldin.atrash, marie-helene.abel, claude.moulin}@utc.fr

Résumé : Beaucoup d'efforts ont été faits durant les deux dernières décennies pour gérer les connaissances dans les organisations, en particulier les connaissances tacites, toujours difficile à transférer contrairement aux connaissances explicites. La capitalisation des expertises personnelles dans les organisations joue un grand rôle dans l'apprentissage organisationnel. Dans cette article, nous proposons une plateforme web qui facilite l'organisation des connaissances dans les organisations. Nous présentons comment les fonctionnalités de cette plateforme permettent de facilement partager les ressources d'information entre les membres d'une organisation. En particulier, nous insistons sur les ressources sociales, issues d'outils collaboratifs comme le chat, le forum ou le wiki, et sur les ressources à finalité sociale comme les notes et les annotations.

Mots-clés : Plateforme Web, Apprentissage organisationnel, Partage des connaissances

1 Introduction

L'apprentissage organisationnel est une propriété émergente de la circulation des idées et de la diffusion des pratiques des membres d'une organisation. Il favorise la compétitivité, la capacité d'innovation et l'efficacité des organisations. Il n'y a pas d'apprentissage organisationnel sans apprentissage individuel (Houdoy, 2000). Cependant l'apprentissage organisationnel représente plus que la somme des apprentissages individuels (Nevis *et al.*, 1995). (Duncan, 1979) précise que afin de faciliter le partage des connaissances dans l'organisation, la connaissance doit être communicable et intégrable. Cela signifie qu'elle doit être représentée d'une manière compréhensible et distribuable et enregistrée dans une mémoire organisationnelle accessible et cohérente. Afin de gérer l'ensemble des ressources hétérogènes circulant dans une organisation, nous proposons la plateforme web MEMORAe (MEMoire ORGanisationnelle appliqué au e-learning). Dans cet article, nous présentons cette plateforme (section 2) et nous montrons comment cette plateforme facilite le partage et l'échange d'information au sein des organisations.

2 La plateforme Web MEMORAe

MEMORAe est une plateforme Web qui exploite la puissance des nouvelles technologies support à la collaboration (technologies web 2.0, etc.). La plateforme est basée sur le modèle sémantique MEMORAe-core 2 (Deparis *et al.*, 2014). Ce modèle considère l'organisation comme un ensemble dynamique de groupes qui peuvent partager des ressources d'information. Chaque membre d'une organisation peut appartenir à plusieurs groupes et des groupes particuliers peuvent se former à l'occasion de nouvelles tâches ou de nouveaux projets. Un utilisateur peut créer un groupe et inviter d'autres membres à y participer. Chaque groupe possède un espace de partage où sont conservées les ressources de ce groupe.

Les ressources sont indexées par un ou plusieurs concepts d'une ontologie qui représente un référentiel métier (ontologie d'application). Cette ontologie est spécifique pour chaque organisation utilisatrice de la plateforme et représente les termes sur lesquels les collaborateurs partagent et échangent des ressources. Cette ontologie est représentée par une carte sémantique qui se situe au milieu de l'interface Web représentée sur la figure 1. Le concept focus est le nœud de la carte qui intéresse un utilisateur à un moment donné. Il est alors déplacé en position centrale de l'écran.

Les espaces de partage peuvent être visualisés en parallèle, ce qui facilite le transfert des ressources d'un espace à un autre via la technique du glisser-déposer. Dans ce cas, une ressource n'est pas dupliquée dans un autre espace ; elle est simplement rendue accessible dans deux espaces différents.

La plateforme intègre un moteur facilitant la recherche de ressources d'information. Celle-ci peut se faire au niveau du contenu d'une ressource ou au niveau de ses métadonnées (auteur, date, taille, etc.). La navigation dans la carte sémantique permet quant à elle une recherche de ressources à partir des concepts sur lesquels elles sont indexées.

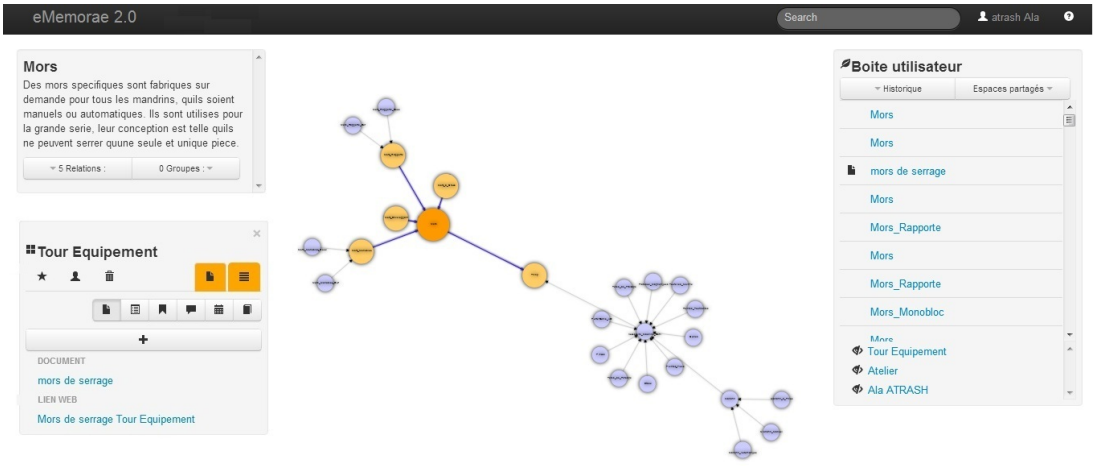


FIGURE 1 – L'interface Web de la plateforme MEMORAe

3 Les ressources d'information

La plateforme MEMORAe intègre trois types principaux de ressources :

– Les ressources documentaires : ce sont des ressources textuelles ou multi-media enregistrées sous forme de fichier (texte, image, vidéo, son, etc.). Ce type est utilisé pour l'échange de rapports, de représentations graphiques, etc.

– Les ressources issues d'un processus social : ce sont les ressources issues de chat, ou contenues dans les forums et les wikis. Les membres d'un groupe peuvent discuter sur un sujet particulier et puis partager et indexer leur discussion pour la consulter plus tard si nécessaire. Ils peuvent également contribuer dans les forums ou les wikis. Ce type de ressource est construit collaborativement.

- Les ressources à finalité sociale (deux types de ressource) :
  - Les notes : une note est la transcription rapide d'une idée, d'une référence, etc. Les notes sont simples à créer et à organiser dans la plateforme MEMORAe. Un cluster de notes est une boîte qui peut contenir des notes et même des clusters de notes. Le partage de notes permet par exemple de réagir rapidement à des idées ou de connaître des avis sur des sujets particuliers.
  - Les annotations : ce sont des traces de lecteurs vis-à-vis d'un document. Elles reflètent les avis des annotateurs et peuvent être utilisées pour expliciter des connaissances. Le lecteur d'un document déjà annoté découvre les avis des annotateurs ce qui peut l'aider à comprendre le document ou l'inviter à approfondir un passage. En utilisant l'outil d'annotation sur l'interface de MEMORAe, un utilisateur peut écrire des annotations directement sur un document. L'utilisateur sélectionne le texte qu'il veut annoter et ajoute une annotation sur le texte sélectionné. Chaque annotation possède un type (commentaire, référence, explication ou question). Les annotations sont accessibles dans les espaces de partage comme ressources à part entière ou en ouvrant le document qu'elles annotent.

4 Conclusion

La plateforme web MEMORAe permet d'organiser les ressources hétérogènes des organisations. Elle facilite le partage d'information entre les membres de l'organisation et elle offre plusieurs chemins d'accès sur une même ressource, ce qui en facilite la récupération à travers de différents espaces de partage. La modélisation des ressources intègre des ressources créées collaborativement ou qui ont vocation à directement inciter la collaboration entre les membres d'une organisation. Les fonctionnalités de la plateforme que nous avons présentées favorisent la circulation de l'information et aident ainsi à l'émergence d'un apprentissage organisationnel.

Références

DEPARIS E., ABEL M.-H., LORTAL G. & MATTIOLI J. (2014). Information management from social and documentary sources in organizations. *Computers in Human Behavior*, **30**, 753 – 759.

DUNCAN R. (1979). Organizational learning : Implications for organizational design. *Research in organizational behavior*, **1**, 75–123.

HOUDOY H. (2000). Réseau d'activités à distance (rad). http ://rad2000.free.fr.

NEVIS E., DIBELLA A. & GOULD J. (1995). Understanding organizations as learning systems. *Sloan Management Review*, p. 73–85.

Infrastructure Web socio-sémantique pour la Veille Collaborative

Jean-Pierre Cahier[1], Mylène Leitzelman[2] et Patrick Brébion[3]

[1] Université de Technologie de Troyes (UTT) - Equipe ICD/Tech-CICO,
cahier@utt.fr
[2] Société Mnémotix, 06210 Mandelieu
mylene.leitzelman@mnemotix.com
[3] Journaliste spécialisé indépendant,
Pbrebion@yahoo.fr

Résumé : L'activité de veille est de plus en plus organisée et gérée en projet collectif, où les équipes de veilleurs doivent multiplier les interactions avec les réseaux d'utilisateurs, en mettant en jeu une variété d'objets (pages Web, articles, projets, équipes, questionnements,...) qu'il s'agit pour les analystes de d'appréhender dans une variété de points de vue et d'activités (analyse qualitative ancrée dans les documents, le catalogage, la visualisation des réseaux sociaux associés aux thèmes du domaine, l'annotation partagée, la co-écriture de rapports, etc.). Nous proposons dans le domaine de la Veille d'expérimenter et évaluer sur plusieurs terrains l'apport d'une approche innovante d'ingénierie de connaissances collaborative utilisant une infrastructure "Web socio sémantique", réunissant notamment les Webmarks (Delaforge *et al.*, 2012) et une série d'outils gérant des points de vue multiples (Cahier *et al.*, 2013).

Mots-clés : web socio-sémantique, travail collaboratif, Webmarks, points de vue, veille.

En accord avec (Weick, 1995) (Baumard, 1991)(Bulinge, 2006), nous pensons (Leitzelman, 2010) que le cycle classique de la veille formalisé par une approche rationnalisée et itérative de l'information a ses lacunes. L'analyse du processus de veille est restée longtemps centrée sur le rapport de l'individu à l'information, or le web 2.0 et l'adoption massive des technologies sociales, virales et mobiles dans la sphère privée puis dans la sphère professionnelle, ont modifié ces rapports de l'individu au groupe dans le monde du travail.

La veille demande une approche collaborative, que nous proposons de baser sur un ensemble d'outils, partageant une infrastructure commune aux standards du Web. Ces outils supposent des modèles de connaissances et de ressources adaptés au travail de groupe, ce qui les place dans le champ du Web socio-sémantique pour assurer la jonction avec les techniques de Web social (modèles d'utilisateurs...) et de Web sémantique (modèles de concepts.) Notre hypothèse, que nous avons commencé à mettre à l'épreuve sur plusieurs terrains, est que cet ensemble d'outils d'ingénierie de connaissances collaborative, est de nature à faciliter grandement l'activité collective de veille, en permettant de travailler plus rapidement avec une qualité de veille accrue sur des champs informationnels plus complexes.

1 La veille comme activité collaborative

Les activités de veille 2.0 que nous considérons ne consistent pas à étudier un corpus de documents primaires figé au départ, mais au contraire à le constituer de façon incrémentale et permanente en cherchant aussi à enquêter au contact de membres des communautés

concernées et en confrontation avec les débats qui les traversent. Ce qui implique de laisser « la main » aux veilleurs sur la définition des sources et des corpus en temps continu.

Répondre à cette problématique en prenant en compte des volumes de documents à surveiller se traduit par l'utilisation d'outils support facilitant l'identification des données, notamment, le signal faible, sans automatiser les étapes d'affectation de thèmes et de catégories. Cette intervention humaine autorise l'enrichissement dynamique des éléments pour ajuster en continu le rapport entre le bruit et le silence. Elle facilite également l'identification du signal faible. Les objets d'intérêt de veilleurs peuvent être détectés ou émerger dans des lectures de documents mais aussi dans des interactions entre veilleurs (e.g. des tags, des recommandations de lectures, des co-écritures de notices intermédiaires, des évaluations de réputation, des chats…) ou entre veilleurs et acteurs du domaine..

Pour dépasser les limites de l'approche classique, un système de veille doit permettre aux acteurs de réajuster facilement les périmètres et les corpus et de mettre en concurrence et en correspondance plusieurs vues sur les tendances latentes. Les étapes suivantes d'enrichissement et d'analyse se doivent également d'être basées sur la même logique pour profiter de cette dynamique et de cette approche multipoints de vue. Des corpus ainsi que des jeux de catégories d'analyse évolutifs et pluriels, seront ainsi de nature à refléter l'existence de plusieurs discours voire de controverses.

2 Infrastructure " Web socio-sémantique " proposée

L'architecture fonctionnelle de l'infrastructure d'outils que nous proposons d'expérimenter, sans être attachée à un processus ni à une méthode de veille uniques, considère la veille comme l'activité collaborative riche décrite précédemment. L'approche technologique suivie consiste à rendre interopérables et utiliser grâce à une plate-forme en Services Web REST une série d'outils de logiciel libre issus des recherches de nos équipes (dont certains ont été développé ou intégrés par les auteurs) tous ayant fait l'objet par ailleurs de publications.

Un premier enjeu est la cons**t**itution d'un paysage thématique partagé d'un domaine vaste, incluant la délimitation et la structuration graduelles du domaine cible. Un premier aspect de la structuration concerne l'identification progressive des items et des corpus pertinents. Ici cette activité utilise les outils Argos et Agoræ (Zaher et al., 2006).

TABLE 1 – *Principales fonctions utilisées dans les expériences projetées et outils correspondants.*

Principales fonctionnalités mentionnées	**Références des outils et/ou démos**
Bookmarking social et sémantique	Plateforme Webmarks (http://mnemotix.com)
Chaîne d'annotation sémantique	Plateforme Webmarks (http://mnemotix.com)
Analyse des réseaux d'acteurs	Plateforme Webmarks / module SemSNA (http://mnemotix.com)
Middleware sémantique (service JSON)	Mnémokit (http://mnemotix.com)
Cataloguage multi-points de vue	http://hypertopic.org/Knowledge_management.html
Analyse qualitative ancrée dans les documents	http://hypertopic.org/Texts_analysis.html
Analyse des co-occurences items/thèmes	https://github.com/Hypertopic/Porphyry/wiki
Middleware socio-sémantique Hypertopic (REST)	https://github.com/Hypertopic/Protocol/wiki

Le modèle Hypertopic V2 (Zhou, 2006) sous-jacent permet de définir plusieurs types d'items et de les taguer selon plusieurs points de vue, qui peuvent être des opinions, des dimensions d'analyse partagées, ou des catégories importées ou construites par l'analyste avec d'autres outils de la plate-forme proposée. Le réseau de thème et d'items aussi être exploité à l'aide d'un autre outil (Porphyry) procurant alors des fonctions supplémentaires de visualisation fine des co-occurrences entre thèmes, rendant alors visible le réseau de relations entre des thèmes utilisés pour décrire un même item, y compris si ces thèmes appartiennent à des points de vue différents. Cassandre, qui est un outil d'analyse qualitative permettant le travail de thématisation et de catégorisation "manuelle" procure par ailleurs aux veilleurs une

aide semi-automatisée en faisant apparaitre, si besoin, des comptages lexicométriques des mots ou groupes spécifiques ou répétés (ou au contraire rares) (Lejeune & Bénel, 2012) pour détecter des régularités ou des singularités dans une lecture de survol. Grâce à l'usage conjugué de l'outil LaSuli (Bénel et al., 2010), organisé tout comme Cassandre par le modèle Hypertopic, les fragments peuvent être tagués selon des points de vue et reliés ainsi à des catégories d'analyse sous contrôle total du lecteur, l'analyste pouvant faire apparaitre une catégorie avant de lui donner un nom, ce qui est classique dans une démarche d'enquête.

Une fonctionnalité de wiki sémantique, incluse également dans la plate-forme proposée, a été développée et déployée dans le but double d'assurer l'édition collaborative de documents d'une part, et en corollaire, de capturer des traces d'activités permettant d'inférer des relations sociales au sein des membres de la communauté de veille testée (Buffa et al., 2012). Par l'intermédiaire de plugin, il est possible d'insérer des visualisations graphiques proposant plusieurs vues de l'activité des membres du réseau social des veilleurs (ex : le graphe social égo-centré personne / personnes / tags partageant les mêmes concepts manipulés, une timeline des tags les plus cités sur une ligne de temps, etc.). Des modèles et des méthodes (Erétéo 2011) permettent de représenter les acteurs, leurs relations (FOAF, RELATIONSHIP), leurs activités en ligne (SIOC), et de structurer les concepts qu'ils manipulent (SKOS).

Plusieurs expériences ont déjà pu être effectuées ou sont en cours sur l'infrastructure proposée, avec des collectifs souvent pluridisciplinaires atteignant jusqu'à 20 veilleurs combinant plusieurs des services évoqués. Notre objectif est de continuer la mise à l'épreuve de cette boite à outils de veille collaborative, pour vérifier cette approche est de nature à faciliter l'activité collective de veille.

Références

BENEL, A., LEJEUNE, C., ZHOU, C., Éloge de l'hétérogénéité des structures d'analyse de textes. Document numérique, RSTI 13(2), 41-56. Hermès-Lavoisier, 2010.

BAUMARD, P. (1991). Stratégie et surveillance des environnements concurrentiels. Masson.

BUFFA M., HUSSON G., DELAFORGE, N.. SweetDeki : le wiki sémantique couteau suisse du réseau social ISICIL.EGC, volume RNTI-E-23 of Revue des Nouvelles Technologies de l'Information, page 387-398. Hermann-Éditions, (2012)

BUFFA M., DELAFORGE N., ERETEO G., GANDON F., GIBOIN A. AND LIMPENS F., "ISICIL: Semantics and Social Networks for Business Intelligence", conference SOFSEM 2013, 39th Int.Conference on Current Trends in Theory and Practice of Computer Science. January 26-31, 2013 Spindler Mlýn, Czech Republic.

BULINGE, F. (2006). Le cycle du renseignement : analyse critique d'un modèle empirique. Market Management , pp 36 à 52.

CAHIER J.-P.,BENEL,A. SALEMBIER, P.,(2013).Towards a "non-disposable" software infrastructure for participation. Interaction Design and Architecture(s) Journal (IxD&A) 18, 68–83. Univ. Roma II.

DELAFORGE N.,GANDON F.,Webmarks: Le marquage d'intérêt sur le Web de données, 12e Conf. Extraction & Gestion des Connaissances(EGC), Bordeaux, France, 2012

ERÉTÉO, G. (2011). Semantic social Network Analysis. PhD Thesis.

LEJEUNE CH., BENEL A., Lexicométrie pour l'analyse qualitative : Pourquoi et comment résoudre le paradoxe.Actes des 11e journées internationales d'analyse statistique de données textuelles (JADT), Lexicometrica.2012.

LEITZELMAN, M. (2010, janv). La veille 2.0 : Outiller les interactions sociales au sein du processus de veille. Les Cahiers du numérique : Du web 2.0 au concept 2.0 , Volume 6, p.200.

WEICK, K. (1995). Sensemaking in organizations. Thousand Oaks: Sage Publications.

ZAHER L'H., CAHIER J.-P., ZACKLAD M., The Agoræ/Hypertopic approach, Proceedings of the workshop on Indexing and Knowledge in Human Sciences, Nantes, June 26-28, 2006. 66-70.. 2006.

ZHOU C., LEJEUNE CH. AND BÉNEL A., Towards a standard protocol for community-driven organizations of knowledge, in Proceedings of the 13th International Conference on Concurrent Engineering (ISPE CE'06), (2006) IOS Press.

Suis-je celui que je prétends être ?

Diyé Dia[1,2], Olivier Coupelon[1], Yannick Loiseau[2], Olivier Raynaud[2]

[1] ALMERYS, solution santé d'Orange Business Services, Clermont-Ferrand, France
diye.dia, olivier.coupelon@almerys.com

[2] LIMOS, Informatique, Modélisation et Optimisation des Systèmes, Clermont-Ferrand, France
yannick.loiseau@univ-bpclermont.fr
raynaud@isima.fr

Résumé : L'usurpation d'identité est une fraude génératrice de grande méfiance des internautes envers l'utilisation des services numériques en ligne. La mise en place d'un système d'authentification implicite, c'est-à-dire basée sur l'étude du comportement de l'internaute, est un moyen original pour lutter contre cette fraude, pour restaurer la confiance et ainsi favoriser l'usage des services en lignes. Dans notre étude, l'authentification implicite se présente comme un élément de sécurité complémentaire aux éléments de sécurité traditionnels. Le rôle de notre système d'authentification est de détecter le plus tôt possible qu'un internaute n'est pas celui qu'il prétend être et/ou de valider le plus longtemps possible son identité. Plus précisément, ce système automatique peut être appelé à la demande - mode ponctuel - pour permettre l'accès à une fonctionnalité plus critique par exemple ou en mode continu pour élever le niveau de sécurité global de la plateforme d'accès. Le principe théorique du système consiste à générer des signatures pour chaque utilisateur à partir de l'historique de son comportement et de comparer ces signature à la trace locale pour valider ou non son identité.

Mots-clés : Comportement utilisateur, identité, authentification implicite, sécurité, confiance

1 Introduction

L'entreprise Almerys, porteuse du projet, déploie un espace de vie numérique composé d'un ensemble de services. Parmi ces service nous trouvons par exemple un coffre fort numérique, un accès à des communautés virtuelles et à des services d'e-commerce [1]. Pour accéder à cet espace et ainsi utiliser l'ensemble des fonctionnalités de la plateforme, un internaute doit se connecter avec un moyen d'authentification classique (login/mot de passe ou carte à puce/code PIN). Seulement, ce déploiement se fait dans un contexte de crise de confiance généralisée envers les systèmes en lignes. Cette crise étant en grande partie liée à la multiplication des vols d'identité sur Internet (120000 victimes d'usurpation d'identité par an en France [2]). Pour lutter contre ce type de fraude, il apparait nécessaire de consolider les systèmes de sécurité traditionnels mais sans détériorer le niveau d'usage des services et tout en respectant la vie privée des internautes. On appelle authentification implicite une authentification basée sur l'étude du comportement de l'internaute. A titre d'exemple, l'adresse IP d'une machine ou la géolocalisation d'un utilisateur peuvent être interprétées comme une forme de signature. Ainsi la mise en place d'un tel système, accepté par l'utilisateur et ensuite transparent pour lui, est un moyen original pour lutter contre les fraudes, pour restaurer la confiance et ainsi favoriser l'usage des services en lignes.
Dans notre étude, l'authentification implicite se présente comme un élément de sécurité complémentaire aux éléments de sécurité traditionnels. Le rôle de notre système est de détecter le plus tôt possible qu'un internaute n'est pas celui qu'il prétend être et/ou de valider le plus longtemps possible son identité. Ce système automatique peut être appelé à la demande - mode ponctuel -

1. https ://www.ebeeoffice.ca/ebee-home/public
2. http ://www.lepopulaire.fr/limousin/actualite/2013/02/25

pour permettre l'accès à une fonctionnalité plus critique par exemple ou en mode continu pour élever le niveau de sécurité global de la plateforme d'accès.

Dans la section suivante nous donnons un aperçu de l'état de l'art, et une approche à notre problème est décrite dans la section 3.

2 État de l'art

L'authentification implicite sur les téléphones mobiles a été étudiée par (Shi *et al.*, 2011) en se basant sur des caractéristiques propres aux smartphones tels que les appels, les sms, la navigation entre les applications du smartphone et la localisation. Les expériences qu'ils ont faites à partir des données de 50 utilisateurs sur 12 jours sont prometteurs malgré l'insuffisance des données. De même, l'étude de (Abramson & Aha, 2013) applique son modèle aux données de 10 utilisateurs sur 1 mois. Les auteurs utilisent des données de navigation web multi-sites pour faire de l'authentification. Ils se basent sur la date/heure des rêquetes et l'url visitée. Une étape de pré-traitement leur permet d'extraire d'autres types de données pour leur étude tels que le genre de la page visitée. Le travail de (Yang, 2010) permet de faire de l'identification en se basant sur des données de navigation web multi-sites également. Elle utilise comme éléments caractéristiques du comportement, le nom du site visité, le nombre de pages vues, l'heure de démarrage d'une session et la durée d'une session. Pour construire le profil de l'utilisateur, son comportement passé est étudié à l'aide de modèles comportementaux. La probabilité qu'il soit celui qu'il prétend être est calculé en comparant le comportement récent, lors d'une session par exemple, avec le profil de l'utilisateur. Les modèles comportementaux sont des modèles statistiques ou des modèles qui s'appuient sur des algorithmes de fouille de données. La classification bayésienne est utilisée dans (Ullah *et al.*, 2011) pour construire le profil des utilisateurs de streaming vidéo afin de prédire l'identité de ces derniers mais les résultats peuvent être améliorer. L'étude de (Abramson & Aha, 2013) utilise la machine à vecteurs de support de LibSVM sous le logiciel Weka pour construire le profil des utilisateurs. Cette étude conclut que les caractéristiques utilisés ne sont pas suffisants pour authentifier ou distinguer les utilisateurs. (Yang, 2010) utilise des calculs de fréquence d'apparition des itemsets construits par l'algorithme Apriori (Agrawal *et al.*, 1996). Sa méthode devient inefficace pour de très grands ensemble de données. Les 300 premières sessions de 2798 utilisateurs suivis sur une année constituent les données de (Yang, 2010). Il est donc nécessaire de bien choisir les éléments qui caractérisent le comportement ainsi que le modèle comportemental permettant de construire le profil.

3 Approche proposée

Une brique d'authentification implicite évalue le comportement de l'utilisateur dans notre plateforme de services numériques. L'authentification implicite peut être continue (l'utilisateur est authentifié à chaque requête demandée) ou ponctuelle (l'utilisateur est authentifié à chaque demande d'accès à une fonctionnalité critique). La criticité de chaque fonctionnalité est évaluée par un expert métier. Cet expert se base sur le niveau de sensibilité de la fonctionnalité. Plus une fonctionnalité est sensible, plus l'impact d'une attaque sur cette fonctionnalité est importante. Pour accéder à une fonctionnalité critique, la probabilité qu'il soit celui qu'il prétend être doit être maximum. Pour la première utilisation de l'authentification implicite, il sera nécessaire

de demander à l'utilisateur s"il souhaite activer le module d'authentification implicite afin de respecter les recommandations de la CNIL [3]. Nous construisons notre profil utilisateur à partir de données de connexion et de navigation sur une plateforme de services. Nous avons moins de données que si nous avions utilisé l'ensemble des données de navigation à travers le web, mais l'étude de (Shi *et al.*, 2011) montre que nous pouvons construire des profils prometteurs avec peu de données. Les fonctionnalités sont regroupées par service sur la plateforme de services. Les éléments qui caractérisent le comportement de l'utilisateur sont le nom de la fonctionnalité demandée, date/heure de la demande, le début de la session, la durée de la session, le service auquel appartient la fonctionnalité demandée et l'adresse IP (anonyme). Nous regroupons nos instances par utilisateur et par moment de la journée. Nous proposons trois moments de la journée : "Matin", "Après-midi" et "Soir". Dans notre base d'apprentissage, nous cherchons les itemsets fréquents avec l'algorithme Apriori pour chaque utilisateur et pour chaque moment de la journée. Un itemset de taille et de support maximal est considéré comme un profil. Un utilisateur a un profil pour chaque moment de la journée. Pour évaluer notre approche, nous construisons une matrice de confusion en utilisant notre base de test. La matrice est remplie en regardant si le profil est inclus dans les instances de la base de test. Nous testerons notre approche sur les logs de 30 utilisateurs sur 15 jours.

4 Conclusion

Nous sélectionnons des éléments caractérisant le comportement qui sont spécifiques à notre contexte. Notre modèle comportemental est intuitif et simple, par rapport aux modèles cités dans la littérature. Une comparaison de nos résultats avec ceux de (Yang, 2010) permettra d'évaluer la qualité de notre approche. Nous allons prendre en compte nos contraintes les plus importantes à savoir le respect de la vie privée des utilisateurs ainsi que la détection très rapide d'un imposteur.

Références

ABRAMSON M. & AHA D. W. (2013). User authentication fromweb browsing behavior. In *Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference*, p. 268–273.

AGRAWAL R., MANNILA H., SRIKANT R., TOIVONEN H. & VERKAMO A. (1996). Fast discovery of association rules. In *Advances in knowledge discovery and data mining 12 (1), AAAI Press*, p. 307–328.

SHI E., NIU Y., JAKOBSSON M. & CHOW R. (2011). Implicit authentication through learning user behavior. In *M. Burmester et al. (Eds.) : ISC 2010, LNCS 6531, Springer-Verlag Berlin Heidelberg*, p. 99–113.

STOCKINGER T. (2011). Implicit authentication on mobile devices. In *the Media Informatics Advanced Seminar on Ubiquitous Computing*.

ULLAH I., BONNET G., DOYEN G. & GAÏTI D. (2011). Un classifieur du comportement des utilisateurs dans les applications pair-à-pair de streaming vidéo. In *CFIP 2011 - Colloque Francophone sur l'Ingénierie des Protocoles*.

YANG Y. C. (2010). Web user behavioral profiling for user identification. In *Decision Support Systems, Elsevier*, number 49, p. 261–271.

3. http ://www.cnil.fr/

Agrégation pour la réparation de liens

Léa Guizol

LIRMM, UNIVERSITÉ DE MONTPELLIER II, CNRS, Montpellier, France
INRIA, Sophia-Antipolis, France
lea.guizol@lirmm.fr

Résumé :
Nous développons un système d'aide à la décision dans le contexte du projet Qualinca afin d'aider les bibliothécaires lors de l'ajout d'une nouvelle description de document. Nous discutons une méthode de validation des liens.
Mots-clés : système d'aide à la décision, agrégation de critères, entité résolution

1 Introduction

L'Abes, Agence bibliographique de l'enseignement supérieur, gère le Sudoc [1] (Système Universitaire de Documentation, une grande base bibliographique) depuis 2001. Le Sudoc contient environ 10 millions de descriptions de documents, ou **notices bibliographiques**, et 2,4 millions de **notices d'autorités**, des descriptions d'entités (lieux, personnes, évènements, ect.) utiles pour décrire les documents. Les notices bibliographiques sont reliées par des **liens** aux notices d'autorités identifiant des entités reliées au document décrit.

Lorsqu'un(e) bibliothécaire souhaite ajouter la description d'un livre au Sudoc, il créé une nouvelle notice bibliographique. Il renseigne les attributs de la nouvelle notice bibliographique (titres, ISBN, nombre de pages...) d'après le livre qu'il a entre les mains. Les **contributeurs** [2] sont représentés par des liens vers les notices d'autorité représentant ces personnes. Par conséquent, le(la) bibliothécaire doit chercher dans le Sudoc chaque notice d'autorité représentant l'un des contributeurs, à l'aide d'une fonction recherchant les notices d'autorités susceptibles de décrire une personne ayant une **appellation** (nom et prénom) donnée. Si il y en a plusieurs (homonymes ou orthographe proche), le(la) bibliothécaire doit décider quelle notice d'autorité représente le mieux la personne. Il(elle) peut aussi en créer une nouvelle pour représenter la personne si aucune ne convient.

Les notices d'autorités sont pauvres en informations. Par conséquent, pour décider si l'une d'elles représente la personne souhaitée, le(la) bibliothécaire doit regarder les informations contenues dans les notices bibliographiques liées à la notice d'autorité. Les erreurs de liage entre notices bibliographiques et notices d'autorité présentes dans le Sudoc favorisent donc l'ajout de nouvelles erreurs de liage.

Dans le but d'améliorer la qualité des données du Sudoc, un travail préliminaire a été présenté dans [1], où une méthodologie générale pour un système d'aide à la décision a été présentée afin de réparer les liens dans une base de connaissances bibliographiques comme celle du Sudoc. La méthode est basée sur le partitionnement des **autorités contextuelles** (objets représentant les

1. http://en.abes.fr/Sudoc/The-Sudoc-catalog
2. Personnes ayant contribué à la réalisation du document.

notices bibliographiques du point de vue d'un contributeur particulier) en fonction de **critères**. La méthode générale consiste en :

1. L'expert(e) entre une appellation A. Le système renvoie un ensemble de notices d'autorités du Sudoc susceptibles de représenter la personne désignée par l'appellation. Chaque notice bibliographique liée à une notice d'autorité sélectionnée est aussi sélectionnée.
2. Une autorité contextuelle est construite pour chaque lien entre une notice d'autorité et une notice bibliographique sélectionnées. Une autorité contextuelle correspond intuitivement à une personne, dans le contexte d'un document auquel elle a contribué.
3. Cet ensemble d'autorités contextuelles constitue le **sous-ensemble du Sudoc de l'appellation A**, noté $ses(A)$. $ses(A)$ est partitionné selon une méthode de partitionnement et un ensemble de critères. Les critères utilisés retournent des valeurs de comparaison symboliques et non pas numériques. Le but de cette étape est d'obtenir les partitions ayant le plus de "sens" selon les critères. Les partitions obtenues peuvent être comparées à la partition initiale (l'unique partition telles que toute et seulement les notices contextuelles issues d'une même notice d'autorité du Sudoc soient dans une même classe) afin de détecter des erreurs de liage dans le Sudoc et d'éventuellement les réparer.

Après avoir exploré comment les liens sont répartis dans le Sudoc et les critères implémentés (Section 2), on présente dans la section 3 un example montrant les limites de l'existant.

2 Données du Sudoc et critères

Nous avons compté le nombre de notices bibliographiques liées à chaque notice d'autorité du Sudoc représentant une personne afin d'observer la répartition des liens. Les liens sont très inéquitablement répartis entre les notices d'autorités :

– 1520285 notices d'autorités sont liées au moins 1 fois ;
– 972 notices d'autorités sont liées au moins 250 fois ;
– 113 notices d'autorités sont liées au moins 1001 fois ;

Les critères de partitions implémentés actuellement sont *domaine*, *date*, *titre*, *appellation*, *contributeurs* et *langue*. Le domaine de publication est représenté dans le Sudoc par une liste de codes de domaines. La distance entre deux codes de domaines et l'agrégation de ces distances ont été fournies par les experts. Les dates de publications sont comparées par rapport aux intervalles de temps entre elles. On utilise une distance de Levenstein adaptée pour comparer les titres. Le critère *contributeurs* donne une valeur de rapprochement en fonction du nombre de contributeurs en commun (excepté celui désigné par l'appellation). Le critère *appellation* est basée sur une fonction de comparaison fournie par les experts qui compare deux appellations (nom et prénom). Le critère *langue* donne une valeur de comparaison d'éloignement si les langues sont distinctes et qu'aucune n'est l'Anglais.

3 Approche et discussion

On considère le sous-ensemble du Sudoc représenté dans le tableau 1. On considère l'ensemble des autorités contextuelles liées à l'appellation "Sam, Harris" ($\{3,4,5,6,7,8\}$). La partition validée de façon experte est $\{\{5\},\{3,4\},\{7\},\{6\},\{8\}\}$. L'attribut "domaine" est une liste

id	titre	date	domaines	[...]	appellations
1	Le banquet	1868			“Platon”
2	Le banquet	2007			“Platon”
3	Letter to a Christian nation		[320,200]		“Harris, Sam”
4	Surat terbuka untuk bangsa kristen	2008	[200]		“Harris, Sam”
5	The philosophical basis of theism	1883	[100,200,150,100]		“Harris, Samuel”
6	Building pathology	2001	[720,690,690,690]		“Harris, Samuel Y.”
7	Aluminium alloys 2002	2002	[540]		“Harris, Sam J.”
8	Dispositifs GAA en technologie SON	2005	[620,620,530,620]		“Harrison, Samuel”

TABLE 1 – Exemple d’autorité contextuelles réelles

de codes de domaines de publication. Deux objets sont considérés par le critère *domaine* comme *proche* s’ils ont au moins un code en commun, et *eloigne* sinon. Deux objets sont considérés par le critère *date* comme étant *eloigne* si il y a plus de 59 ans entre leurs dates de publication. Donc, l’objet 5 est *eloigne* de tous les autres selon le critère *date*. Cependant, le critère *domaine* considère que les objets 3, 4 et 5 sont *proche* deux à deux parce qu’ils ont le code de domaine 200 (=religion) en commun : 3, 4 et 5 devraient être dans la même classe. Le critère *domaine* considère aussi que les objets 6, 7 et 8 sont deux à deux *eloigne* et *eloigne* des objets 3, 4 et 5. La seule meilleure partition selon les critères *domaine* et *date* est donc $\{\{5,3,4\}, \{7\}, \{6\}, \{8\}\}$. Ce n’est malheureusement pas la meilleure partition selon les experts. Nous affirmons que la raison de ce résultat insatisfaisant est due à la façon dont les valeurs de comparaison sont agrégées par de telles approches : comme des valeurs numériques.

Notre travail concerne deux sémantiques de partitionnement qui ajoutent :

– plusieurs niveaux de valeurs de rapprochement et d’éloignement ;
– pas d’interférence entre les valeurs de rapprochement et d’éloignement (par exemple, une valeur de rapprochement ne peut pas effacer une valeur d’éloignement).

Nous avons proposé deux sémantiques de partitionnement basées sur des critères à valeurs non-numériques. Ces sémantiques de partitionnement répondent aux exigences du projet dans lequel elles s’inscrivent, et en particulier au fait que nous ne souhaitions garder les valeurs de comparaison symboliques des critères le plus possible (par opposition aux techniques de partitionnement qui les numérisent pour les manipuler)[2].

Remerciements Ce travail a été soutenu par l’Agence Nationale de la Recherche (projet ANR-12-CORD-0012). Nous remercions chaleureusement Alain Gutierrez et l’ABES.

Références

[1] CROITORU M., GUIZOL L. & LECLÈRE M. (2012). On Link Validity in Bibliographic Knowledge Bases. In *IPMU’2012 : 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, volume Advances on Computational Intelligence, p. 380–389, Catania, Italie : Springer.

[2] GUIZOL L., CROITORU M. & LECLÈRE M. (2013). Aggregation semantics for link validity. *Proc. of SGAI-AI 2013*, p. 359–372.

Un éditeur de définitions formelles pour les connaissances lexicales de la théorie Sens-Texte

Maxime Lefrançois[1,2], Fabien Gandon[2,1], Alain Giboin[2,1], Romain Gugert

[1] Univ. Nice Sophia Antipolis, CNRS, I3S, UMR 7271, 06900 Sophia Antipolis, France

[2] Inria

{ maxime.lefrancois | fabien.gandon | alain.giboin }@inria.fr
romain.gugert@gmail.com

Résumé : Nous menons une étude en IC appliquée aux connaissances lexicales sémantiques de la Théorie Sens-Texte (TST). Avant de développer un formalisme adapté -le formalisme des Graphes d'Unités-, nous avons dû étendre la conceptualisation des prédicats linguistiques et des définitions lexicographiques dans la TST, puis nous avons voulu valider cette nouvelle conceptualisation auprès des lexicographes du projet RELIEF. Nous présentons donc un prototype d'éditeur qui permet de représenter formellement à l'aide de graphes des définitions lexicographiques. Initialement adaptées aux lexicographes, les Graphes d'Unités et une évolution du prototype d'éditeur présenté pourraient être utilisés dans d'autre contextes de l'IC.

Mots-clés : Sémantique lexicale, définitions lexicographiques, Lexicographie, Théorie Sens-Texte, éditeur de graphes, formalisme des Graphes d'Unités.

1 Introduction

Nous menons une étude en IC appliquée aux connaissances lexicales sémantiques de la Théorie linguistique Sens-Texte (TST) (Mel'čuk *et al.*, 1995). Nous nous intéressons en particulier aux définitions lexicographiques, symbolisées par un réseau sémantique dans la TST.

Afin de faciliter le développement d'un formalisme de représentation des connaissances adapté –le formalisme des Graphes d'Unités (Lefrançois & Gandon, 2013)–, nous avons proposé une extension de la conceptualisation de la TST. Cette conceptualisation est intéressante pour d'autres application en IC, car elle estompe la distinction habituelle concept/relation.

Nous présentons un prototype d'éditeur qui permet de représenter formellement à l'aide de graphes des définitions lexicographiques selon cette conceptualisation étendue, que nous avons fait évaluer par des lexicographes [1] du projet RELIEF [2] (Lux-Pogodalla & Polguère, 2011).

Cet article présente la conceptualisation actuelle des définitions lexicographiques dans le projet RELIEF (§2), l'extension de cette conceptualisation que l'on propose (§3), puis le prototype d'éditeur développé (§4). Nous concluons par son évaluation et les perspectives (§5).

2 Les définitions lexicographiques dans la théorie Sens-Texte

La définition lexicographique d'une unité lexicale L présente de façon formelle le sens dénotationnel de L. Par exemple PEIGNE$_{\mathbf{2A}}$ (le peigne du tisserand) peut être défini par :

peigne$_{\mathbf{2a}}$ de X pour Y = ⁽*Outil de tissage qu'une personne X utilise pour démêler les fibres d'un objet Y*⁾

1. La lexicographie est une science qui a pour sujet d'étude l'édition des dictionnaires
2. RELIEF est un projet d'envergure de l'ATILF, CNRS - http://www.atilf.fr/

Dans le projet RELIEF, l'édition d'une définition lexicographique s'effectue en trois étapes (c.f., fig. 1, gauche). Cependant, Wanner (2003) note qu'il est souhaitable de formaliser d'avantage les définitions lexicographiques, en particulier pour des applications de TALN. Une formalisation en vue est sous la forme de réseaux sémantiques (Mel'čuk *et al.*, 1995).

3 Extension de la conceptualisation

Nous avons introduit un niveau sémantique profond pour conceptualiser les sens. On peut y organiser les prédicats sémantiques en une hiérarchie au sein de laquelle des positions actancielles obligatoires, optionnelles ou interdites, et munies de signatures, sont héritées et potentiellement spécialisées (Lefrançois & Gandon, 2013).

Les définitions lexicographiques sont conceptualisées à ce niveau sous forme de multigraphes étiqueté et orientés, dont la visualisation s'inspire de l'UML.

4 Prototype de définitions lexicographiques formelles

Nous avons proposé un workflow en quatre étapes adapté à la nouvelle conceptualisation (Lefrançois *et al.*, 2013) (c.f., fig. 1, droite). L'éditeur dont nous présenterons une démonstration [3] est une implémentation de ce workflow (Gugert, 2013).

Processus suivi dans le projet RELIEF	Processus à suivre avec le nouvel éditeur
1. Sélection d'une étiquette sémantique dans une hiérarchie (Polguère, 2011)	1. Sélection de la structure actancielle de la sémantique profonde
2. Sélection de la structure actancielle sémantique de surface (Mel'čuk, 2004)	2. Édition de la définition formelle
3. Rédaction de la définition lexicographique en XML (Barque *et al.*, 2010)	3. Sélection de la structure actancielle sémantique de surface
	4. Mise en correspondance des structures actancielles

TABLE 1 – à g., workflow des lexicographes du projet RELIEF ; à dr., workflow de l'éditeur.

Pour l'utilisateur, les opérations d'édition de la définition (c.f., fig. 1b) sont des opérations de manipulation de graphes. Le "glisser-déposer" lui permet de sortir une position actancielle, pour éventuellement se rendre compte que sa signature comporte elle-même des positions actancielles, ou pour fusionner des nœuds. Nous nous sommes inspirés de l'UML, mais nous ne connaissons pas d'éditeur d'UML qui permette de "sortir" un attribut à l'extérieur de sa classe pour ainsi obtenir une association vers une autre classe.

Techniquement, ces opérations sont implémentés en JavaScript au dessus de *mxGraph* [4], qui permet la visualisation et la manipulation de graphes.

5 Évaluation et perspectives

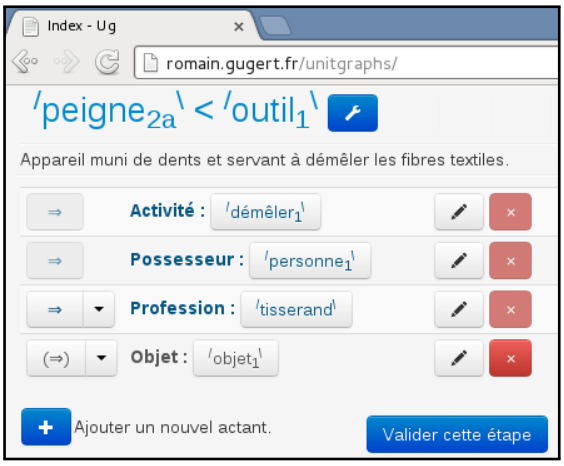
Nous avons évalué l'acceptation du prototype et du processus supportant la nouvelle conceptualisation par la communauté de la TST. Les résultats sont encourageants (Gugert, 2013) et per-

3. Démonstration de l'éditeur - `http://wimmics.inria.fr/doc/video/UnitGraphs/editor1.html`

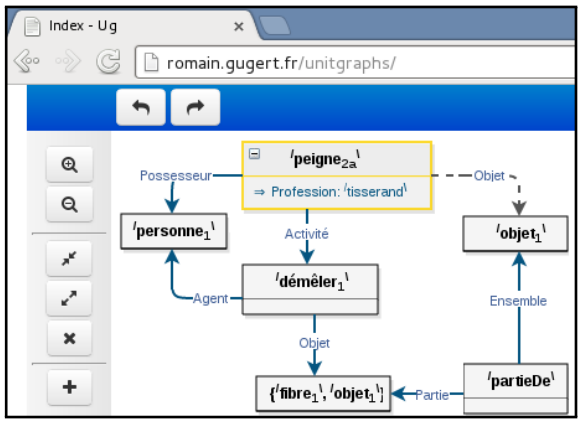
4. mxGraph - visualisation de graphes en JavaScript - `http://www.jgraph.com/mxgraph.html`

mettent d'entrevoir des directions d'amélioration : nous devons en particulier rendre le workflow plus semblable à celui qui est utilisé actuellement.

Initialement adaptées aux lexicographes, les étapes 1 et 2 peuvent être réutilisées dans un autre contexte d'IC. Par ailleurs, une évolution de ce prototype devrait guider l'utilisateur dans l'élaboration de la structure actancielle en même temps qu'il manipule le graphe de définition.



(a) Phase 1 : Sélection de la structure actancielle.



(b) Phase 2 : Édition de la définition formelle par manipulation de graphes.

FIGURE 1 – Captures d'écran du prototype lors de l'édition de la définition de PEIGNE$_{2A}$.

Références

BARQUE L., NASR A. & POLGUÈRE A. (2010). From the Definitions of the 'Trésor de la Langue Française' To a Semantic Database of the French Language. In FRYSKE AKADEMY, Ed., *Proceedings of the XIV Euralex International Congress*, Fryske Akademy, p. 245–252, Leeuwarden, Pays-Bas.

GUGERT R. (2013). Scénarisation d'interactions avec les objets du formalisme des Graphes d'Unités et prototypage d'un éditeur de définitions lexicographiques formelles. Mémoire de Master 2 - UPMC, hal-00860767.

LEFRANÇOIS M. & GANDON F. (2013). Rationale, Concepts, and Current Outcome of the Unit Graphs Framework. In *Proceedings of the 9th International Conference on Recent Advances in Natural Language Processing (RANLP 2013)*, p. 382–388.

LEFRANÇOIS M., GUGERT R., GANDON F. & GIBOIN A. (2013). Application of the Unit Graphs Framework to Lexicographic Definitions in the RELIEF project. In *Proceedings of the 6th International Conference on Meaning-Text Theory (MTT'2013)*, Prague, Czech Republic.

LUX-POGODALLA V. & POLGUÈRE A. (2011). Construction of a French Lexical Network : Methodological Issues. In *Proceedings of the International Workshop on Lexical Resources*, Ljubljana.

MEL'ČUK I. (2004). Actants in semantics and syntax I : Actants in semantics. *Linguistics*, **42**(1), 247–291.

MEL'ČUK I., CLAS A. & POLGUÈRE A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Paris/Louvain-la-Neuve : Duculot.

POLGUÈRE A. (2011). Classification sémantique des lexies fondée sur le paraphrasage. *Cahiers de lexicologie*, **98**, 197–211.

WANNER L. (2003). Definitions of Lexical Meanings : Some Reflections on Purpose and Structure. In *Proceedings of the First international conference on Meaning-Text Theory (MTT'2003)*, p. 16–28.

Un Wiki/IDE pour exploiter le web de données

Pavel Arapov, Michel Buffa et Amel Ben Othmane

Equipe Wimmics, commune aux laboratoires
INRIA et I3S de Sophia Antipolis,
{arapov, buffa, abenothm}@i3s.unice.fr

Résumé : Ce papier décrit WikiNEXT, un moteur de wiki créé pour la rédaction des applications web qui exploitent le web de données directement dans le navigateur web. WikiNEXT est un wiki à la croisée des wikis et des outils de développement en ligne (« web based IDEs[1] »), ce qui fait son originalité. WikiNEXT propose aujourd'hui des fonctionnalités et une interface plutôt orientée vers les développeurs web voulant manipuler des données sémantiques à l'aide des technologies « front end » (JavaScript/HTML5), et offre des moyens pour bénéficier de services proposés par le wiki pour la persistance côté serveur, en fournissant notamment une base de données de graphe compatible RDF/SPARQL 1.1, et une base de données objet.

Mots-clés : wiki, wiki sémantique, IDE basé Web, application Web, Web de données.

1 Introduction

Depuis l'apparition du premier wiki en 1995, créé par Ward Cunningham, (Leuf et Cunningham 2001) de nombreux moteurs de wikis sont apparus, proposant des fonctionnalités communes. MediaWiki est un exemple de moteur de wiki pour faire tourner Wikipedia. Plus tard, les wikis sont devenus très populaires comme des systèmes de gestion de la connaissance. Cette classe de wikis s'appelle les « wikis d'application ». On trouve dans cette catégorie[2] Confluence, Mindtouch, TWiki. Une autre évolution intéressante des moteurs de wikis, issue du monde de la recherche académique, est apparue en 2005 : les « wikis sémantiques ». Ces wikis sémantiques tels que MediaWiki (Krötzsch et Vrandecic 2006), MoKi (Rospocher et al 2009), OntoWiki (Auer et al. 2006) et IkeWiki (Shaffert 2006), ont permis d'étendre l'approche des wikis classiques avec le contexte sémantique tout en préservant la simplicité et l'essence collaborative des wikis. Récemment, avec le support de la technologie Ajax et le développement de JavaScript, les éditeurs de code source pouvant fonctionner dans une page Web n'ont cessé de s'améliorer. Certains éditeurs comme Cloud9[3], Nitrous.IO, jsfiddle.com, et jsbin.com sont très connus par les développeurs.

2 WikiNEXT

WikiNEXT, un mélange entre un IDE basé Web et un wiki classique, permet de développer des applications web qui exploitent le web de données. Plusieurs éditeurs sont fournis : un éditeur WYSIWYG pour des documents classiques, et un éditeur de code pour la partie « application » qui peut être utilisé en mode développement. Le prototype est open source et disponible en ligne[4]. Il inclut aujourd'hui de nombreux exemples et tutoriaux.

1 IDE = Integrated Development Environment.

[2] http://twiki.org, https://atlassian.com/fr/software/confluence, http://mindtouch.com

3 https://c9.io/

[4] http://wikinext.gexsoft.com/

3 Architecture

La Figure 1 décrit l'architecture logicielle de WikiNEXT. Une page WikiNEXT est considérée comme une application web. Ces applications exploitent des ressources externes (i.e., Freebase.com ou DBPedia.fr) ou internes (données stockés dans la base de données). Une application dans WikiNEXT contient du code HTML/CSS/JavaScript. Dans un paradigme MVC, la partie JavaScript de la page formera « la couche métier et le contrôleur », la partie « vue » se fera dans le code HTML / CSS de la page, et la partie « modèle » sera composée de métadonnées RDF. Chaque page WikiNEXT est associée avec un ensemble de métadonnées qui décrivent ses principales caractéristiques : titre, auteur, contributeurs, dernière date de modification, versions, etc. Mais les pages sont aussi des « containers » : elles contiennent également des métadonnées qui auront pu être rajoutées manuellement ou par une application d'annotation. WikiNEXT s'appuie sur les ontologies disponibles sur schema.org pour décrire la structure des pages, les utilisateurs, les applications web ou le contenu des pages. Les métadonnées générées dans les pages de wiki sont stockées dans un triple store intégré à la partie serveur du Wiki lors de la sauvegarde. Nous utilisons pour gérer les triplets RDF une version améliorée de RDFStore, RDFStore-js, un moteur de graphe compatible SPARQL 1.1, écrit en JavaScript. Le contenu traditionnel de la page est stocké sous forme d'objets dans la base de données MongoDB[5]. Nous l'utilisons également comme couche de persistance pour RDFStore-js.

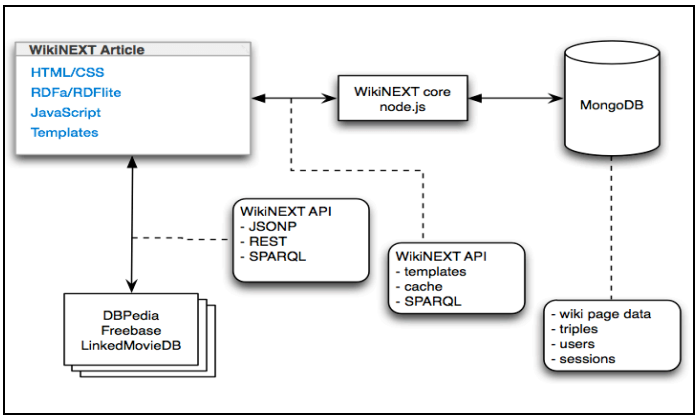


Figure 1: Architecture logicielle de WikiNEXT

4 Démonstration

Dans cette section nous décrivons un scénario d'usage complet : le développement d'une application qui utilise le SPARQL endpoint de DBPedia.org pour récupérer des données RDF concernant des villes (description, population, photos, etc.).

Un formulaire de saisie permet à un utilisateur de saisir le nom des villes qu'il souhaite afficher. Le résultat de la requête est affiché en deux modes : le mode par défaut « sparql On Fly » et en utilisant des pages modèles (templates) basées sur le framework mustache.js (voir la différence entre le deux modes d'affichage dans Figure 2).

Ces résultats sont utilisés pour créer à la volée des pages du wikis –une par ville-, basée sur un modèle de présentation contenant aussi les métadonnées, lui aussi créé dans le wiki comme une page template. Le template gère à la fois l'affichage des données extraites à partir de

[5] http://mongodb.org

DBPedia et les annotations RDFa. Les pages créées sont annotées et les annotations sont sauvegardées dans le triple store RDF du wiki.

Pour démontrer la réutilisation des données dans notre wiki, on a créé une page WikiNEXT qui réalise un « mashup sémantique » construit en requêtant la base de connaissances globale de WikiNEXT sur les villes stockées au lieu de requêter le web de données. Le résultat est affiché dans une carte présentant l'ensemble des villes qui ont été récupérées, avec un résumé et des photos de chacune d'entre elles.

Figure 2: Différence entre l'affichage en mode "sparql On Fly" (à gauche) et en utilisant un template (à droite)

Cette application est disponible en ligne sur le site de WikiNEXT[6], chaque utilisateur peut consulter le code de l'application, le modifier ou la cloner pour avoir sa propre version.

5 Conclusion

Nous avons présenté WikiNEXT qui est un mélange entre un wiki sémantique et un IDE basé Web pour exploiter les données liées. Nous avons également présenté l'architecture globale de notre application. La démonstration que nous proposons consiste à présenter les principes de WikiNEXT pour la rédaction de documents et des applications exploitant le web de données directement dans le navigateur. Nous vous proposons de reproduire en direct, le scénario de la ville décrit dans ce papier, qui montre comment WikiNEXT simplifie la programmation de ces applications, par rapport à des solutions classiques.

Références

Auer, S., Dietzold, S., & Riechert, T. (2006). OntoWiki–A tool for social, semantic collaboration. In The Semantic Web-ISWC 2006 (pp. 736-749). Springer Berlin Heidelberg.

Krötzsch, M., Vrandečić, D., & Völkel, M. (2006). Semantic mediawiki. In The Semantic Web-ISWC 2006 (pp. 935-942). Springer Berlin Heidelberg.

Marco Rospocher, Chiara Ghidini, Viktoria Pammer, Luciano Serafini, Stefanie N. Lindstaedt: MoKi: the Modelling wiKi. SemWiki 200

Schaffert, S. (2006, June). IkeWiki: A semantic wiki for collaborative knowledge management. In Enabling Technologies: Infrastructure for Collaborative Enterprises, 2006. WETICE'06. 15th IEEE International Workshops on (pp. 388-396). IEEE.

[6] http://wikinext.gexsoft.com/wiki/519e04c580194c4178000001

Adnosco : gérez les données que vous diffusez !

Nadia Bennani, Emmanuel Gaude, Előd Egyed-Zsigmond, Philippe Lamarre

Université de Lyon
CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France
firstname.lastname@liris.cnrs.fr

Résumé :
Le nombre de formulaires en ligne a explosé avec le développement du web. C'est le moyen le plus répandu pour acquérir des informations auprès des utilisateurs. Actuellement, ces données sont stockées par les industriels ou les administrations. Les utilisateurs ont un rôle entièrement passif qui les rend dépendants pour la gestion de leurs informations. Nous proposons de démontrer que, convenablement modélisées, la gestion de ces informations par les utilisateurs peut leur être bénéfique aux deux parties : utilisateurs et fournisseurs de services. Cette démonstration a donc pour vocation de présenter un modèle et un prototype basés sur la qualification sémantique de formulaires en ligne qui améliorent la productivité de l'utilisateur.
Mots-clés : Formulaire web, alignement sémantique, complétion.

1 Introduction

De nos jours, la plupart des sites Web archivent les informations personnelles de leurs utilisateurs. De plus, ils disposent de suffisamment d'outils (CRM, fouille de données, etc) pour permettre l'acquisition, le stockage, l'accès, la gestion et l'exploitation des informations obtenues auprès de l'utilisateur. A l'inverse, l'utilisateur final ne dispose d'aucun moyen pour effectuer les mêmes opérations sur ses propres données. Pour illustrer cette lacune, posons nous la question suivante : "quel système permet à un utilisateur de déterminer les informations qu'il a transmises, à quels sites, dans quel contexte ou objectif ?". Sans remettre en cause l'utilité et le droit qu'ont les sites marchands à stocker les données de l'utilisateur, il nous paraît naturel et très complémentaire de donner à ce dernier les moyens de gérer ses propres données. De plus, cela permet d'imaginer d'autres scénarios. Par exemple, suite à un achat en ligne, les informations acquises peuvent être transmises à d'autres applications : *montant* au gestionnaire de comptes bancaires, *date estimée de livraison* à l'agenda. . .. Actuellement, à notre connaissance, ce type d'usage n'existe pas.

Cette approche orientée utilisateur rejoint celle développée par Berkman Center For Internet & Society à l'Université d'Harvard qui propose la notion de *VRM* [3] (Vendor Relationship Management), complémentaire à celle plus connue des *CRM* [5] qui elle, est plus bénéfique aux marchands.

Pour œuvrer dans ce sens, dans cette démonstration, nous présentons *Adnosco*, un outil dédié à la gestion et à l'intégration des données utilisateur. *Adnosco* s'attaque aux problèmes d'acquisition, de stockage, de structuration de données, dans le but de les interroger et de les exploiter. Par manque de place, dans cet article, nous ne discuterons pas les problèmes d'acquisition. L'exploitation des données sera vue sous l'angle restreint de l'assistance à la complétion de formulaires Web, bien qu'en réalité, *Adnosco* permettrait l'exploitation des données utilisateurs dans un contexte plus large.

Une étude récente [2] démontre que la saisie de formulaires est la cause du renoncement à un achat en ligne pour 24% des utilisateurs. Les fonctionnalités d'auto-complétion et de pré

remplissage ont prouvé leur aptitude à augmenter la productivité des utilisateurs dans des environnements professionnels (outils de gestion de stock, de clients…). C'est ce qui a motivé notre choix de la complétion pour démontrer la capacité d'*Adnosco* à gérer et exploiter les données de l'utilisateur avec précision et efficacité. Les solutions actuelles permettent de compléter un champ d'un formulaire uniquement en proposant un ensemble de valeurs issues du même champ dans d'autres instances du même formulaire ou pour des informations très spécifiques (par exemple, le nom, l'adresse, les numéros de téléphone, les cartes de crédits…). La complétion dans *Adnosco* améliore considérablement ces fonctionnalités en proposant pour un champ, toutes les valeurs sémantiquement pertinentes, issues de tous les formulaires préalablement remplis. Nous proposons 2 outils de complétion, l'un syntaxique, l'autre sémantique.

2 Modèle de gestion de données

Adnosco est basé sur un modèle de stockage de données en trois parties (figure 1). La première, *data management*, analyse chaque formulaire, lors de sa soumission, pour acquérir les données transmises dans les champs de celui-ci (ou suite à une demande explicite de l'utilisateur) et stocke les valeurs ainsi obtenues dans une base de données en les reliant à ce formulaire (uri, identifiants, date, types et valeurs des champs). La seconde, *semantic concepts*, définit des *concepts*. Son rôle est de fournir les outils sémantiques, comme les schémas, ontologies, concepts, attributs, types, relations,… servant à caractériser les informations diffusées. La troisième partie, *semantic qualification*, relie les deux premières. Afin de rendre aisé et précis, l'alignement sémantique des formulaires web , nous introduisons la notion de *concept matérialisé* (CM). Un concept matérialisé formalise l'expression d'une occurrence de concept sémantique (par exemple une personne) dans un formulaire (voyageur) en établissant des correspondances entre les propriétés que le concept matérialisé (voyageur) hérite du concept (personne) et les champs du formulaire.

L'assistance syntaxique est relativement limitée car elle ne s'appuie que sur la première partie. L'assistance sémantique utilise les 3 parties et peut donc être bien plus précise sur l'usage des informations collectées.

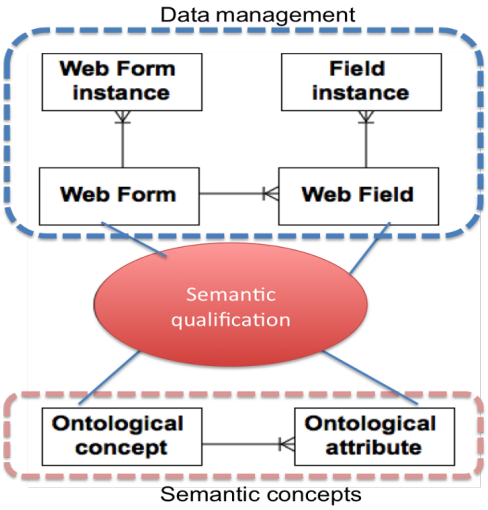


FIG. 1 – Modèle de stockage données d'*Adnosco*

C'est grâce aux concepts matérialisés que la complétion et le pré remplissage des champs sont traités correctement dans le cas où plusieurs occurrences d'un même concept sont présentes dans un même formulaire. Par exemple, un formulaire de voyage peut impliquer plusieurs personnes. Sa qualification sémantique fera apparaître autant de concepts matérialisés '*Personne*' distincts que nécessaire.

Lorsque l'utilisateur veut remplir un champ associé à un concept matérialisé (ex. Voyageur1), le système se concentre sur les champs de ce CM et propose pour complétion des valeurs correspondant aux propriétés d'instances du même concept (ex. Personne) compatibles avec les informations déjà saisies. L'aide au remplissage de formulaires web est donc précise et efficace. Une description plus détaillée d'un scénario illustrant les avantages d'*Adnosco* est décrite dans [1].

3 Conclusion et travaux futurs

Dans cette démonstration, nous proposons *Adnosco*, un système de gestion des données, centré utilisateur. *Adnosco* offre une solution d'acquisition, de stockage et d'interrogation des données transmises à des tiers via des formulaires web. Son service de complétion fournit une aide au remplissage des formulaires grâce à l'exploitation de données déjà transmises. Il met en œuvre des techniques syntaxiques et sémantiques (concepts matérialisés) qui s'appliquent à tout type d'information et qui permettent, par exemple, de traiter correctement des formulaires web contenant plusieurs instances d'un même concept. En résumé, *Adnosco* généralise des solutions comparables [2, 4] et améliore la productivité de l'utilisateur que ce soit dans le cadre d'activités personnelles ou professionnelles. *Adnosco* ne s'oppose pas aux activités de collecte d'informations de type CRM, mais il apporte aux utilisateurs une solution complémentaire rétablissant l'équilibre des rôles dans la gestion de leurs données personnelles. De leur côté, les organismes collecteurs peuvent guider la saisie de l'utilisateur en s'appuyant sur cet outil (qualification sémantique de leurs formulaires) pour améliorer la qualité des informations obtenues.

Références

[1] ADNOSCO (2014). adnosco. http://adnosco.liris.cnrs.fr/doku.php?id=scenario.

[2] DASHLANE (2013). Dashlane. https://www.dashlane.com/download/Dashlane_IFOP_release_2013-03-26_en.pdf.

[3] HAVARD (2007). Vendor management system. http://cyber.law.harvard.edu/.

[4] MIT (2013). openpds. http://openpds.media.mit.edu/.

[5] WIKIPEDIA (2013). Customer management system. http://en.wikipedia.org/wiki/Customer_relationship_management.

Vers une cartographie participative basée sur la communication transversale des acteurs dans les situations de crise

Amina Saoutal[1], Jean-Pierre Cahier[1], Nada Matta[1]

[1] Laboratoire ICD/Tech-CICO, Université de Technologie de Troyes (UTT)
12 rue Marie Curie, 10010- Troyes Cedex, France
{amina.saoutal, jean_pierre.cahier, nada.matta}@utt.fr

Résumé : Nous présentons dans ce poster un modèle de représentation graphique des différentes données et informations utiles à la communication entre les multi-intervenants dans une crise contribuant à une meilleure « awareness » et à des formes participatives dans la gestion de cette crise. Vu la multitude des organisations intervenant dans une crise (professionnels, volontaires…) et les différences de cultures de métier, chaque unité a son objectif et sa priorité, ainsi que sa terminologie pour communiquer les messages. Cela rend plus difficile pour les acteurs de différents métiers présents sur le terrain de communiquer et d'échanger les informations, avec pour conséquence, une insuffisante conscience mutuelle des actions entreprises par les autres acteurs. Notre objectif est d'augmenter cette conscience mutuelle en permettant notamment aux acteurs de participer comme contributeurs d'informations sur une carte géographique, leur offrant une meilleure prise sur la situation et facilitant les activités et la prise de décision dans les meilleurs délais. Cette approche, qui s'appuie sur le modèles IC collaborative du Web socio-sémantique, inclut la représentation sur une carte, la localisation des différents acteurs présents sur le terrain de crise ainsi que les points d'intérêts nécessaires aux activités que les acteurs peuvent ajouter au fur et à mesure de la progression de l'évènement.

Mots-clés : Awareness, communication, gestion de crise, système participatif, ingénierie des connaissances, modélisation, partage d'information.

La gestion de crise mobilise plusieurs acteurs appartenant à différentes organisations et ayant différent objectifs. Parmi les facteurs de réussite de la gestion de crise figure en bonne place la conscience mutuelle (Awareness) des activités des acteurs et de la progression des évènements pendant cette crise. Celle-ci nécessite la communication et le partage d'informations pertinentes et utiles, selon les points de vue des différents acteurs impliqués et présents sur le terrain.

Les intervenants sur une scène de crise rencontrent des problèmes de communication liés à la fois à la transmission d'information, en particulier dans des zones rurales, à la compréhension du message qui dépend de la façon dont le récepteur le perçoit et l'interprète, ainsi qu'aux cultures, priorités, objectif et terminologies de métier qui peuvent différer d'une organisation à une autre. Ces différences entre les unités de secours influencent la communication transversale inter-organisationnelle et font obstacle à la conscience mutuelle.

1 Démarche

Afin de mieux comprendre les problèmes de communication et partage d'informations rencontrés par les différents services et organisations de secours, nous étudions et analysons des cas d'urgence et scénarios dans le cadre d'un projet en Région Champagne-Ardenne. En exploitant les expériences des agents de secours et en modélisant la communication d'information via les acteurs. Nous avons mené des entretiens semi-directifs et débriefing

d'exercices avec les principaux intervenants de secours dans la gestion de crise : les pompiers, des membres de la police et des spécialistes en services d'urgence médicale, dans une première étape de cette recherche nous nous fixons les objectifs suivants 1- Explorer les activités et les challenges rencontrés par les équipes intervenantes. 2- Définir et modéliser comment les acteurs sur le terrain communiquent transversalement au cours de la gestion de crise. 3- Définir les informations utiles à la communication et comment elles sont émises et appréhendées par les acteurs. 4- Analyser les effets de la communication actuelle sur la conscience mutuelle collective et explorer des pistes d'amélioration. Nous avons réalisé les trois premiers objectifs et déterminé les dépendances entre acteurs, information et activités (Saoutal et al. 2014).

Pendant l'analyse, nous avons notamment constaté La communication d'information verticale, cette information est de type *requête*. Ce sont par exemple les demandes d'information, instructions et missions que le commandant des opérations de secours (COS) pour les pompiers, le commandant de gendarmerie et le médecin envoient sur le terrain pour exécution, d'autres informations sont de type *information descriptive* par exemple les information remontées du terrain vers un niveau supérieur pour décrire la reconnaissance réalisée.

Cependant, cette communication d'information reste soit ascendante ou descendante (figure1a), Il est rare qu'on ait affaire à un échange d'information transversal inter-organisationnel, ce qui amène à un manque de conscience sur les activités des autres acteurs, Or certaines activités sont interdépendantes et ne peuvent pas être réalisées sans croiser ou recouper l'information des autres métiers. Une conséquence est la perte du temps ; lorsque tous les services remontent les informations vers le niveau décisionnel, le directeur des opérations de secours (DOS) peut constater que ces informations sont différentes d'une organisation à une autre, voire contradictoires, et dans ce cas demander de revenir sur le terrain et vérifier ces information. Vu que chaque unité à ses propres cultures et terminologies, l'interopérabilité des systèmes et l'interprétation de messages sont davantage des problèmes rencontrés par les différents acteurs. Cela justifie d'utiliser une approche d'IC collaborative de Web socio-sémantique (Zhou et al., 2006) qui permet de caractériser les items de la situation de crise selon des thèmes relevant de plusieurs points de vue d'acteurs, se traduisant par exemple sur la carte de la crise par des systèmes d'icones relevant de ces multiples points de vue (Ma et al. 2012).

A partir des résultats de cette partie études, nous allons concentrer la recherche sur les outils de communication inter-organisationnelle. Notre approche s'appuie sur les approches et modèles du Web socio-sémantique (Zhou et al, 2006) et notamment leur application dans les applications territoriales de façon à encourager la communication transversale des agents de différents métiers sur le terrain en proposant un système basé sur une carte géographique du site de la crise (figure1b). Cette approche va leur permettre de partager des données et informations utiles en prenant en considération les problèmes évoqués précédemment. Ces données et informations pourraient être présentées sous forme de:

- Icônes : dans notre approche, tout en reflétant leurs points de vue métier (Ma et al, 2012) les icones peuvent être directement positionnées par les acteurs depuis des dispositifs mobiles afin de représenter les points d'intérêts sur la carte, comme la zone toxique, les agents des différents métiers ainsi que leurs localisation, l'accessibilité, etc.
- Photos : Afin d'éviter toute interprétation, les photos pourraient être pertinentes pour gagner du temps dans la description d'une situation ou de l'état d'une victime.
- Texte : pour présenter toute donnée exacte, comme le taux de toxication, les produits chimiques, le nombre de victime, la surface contaminée, ainsi que les actions etc.

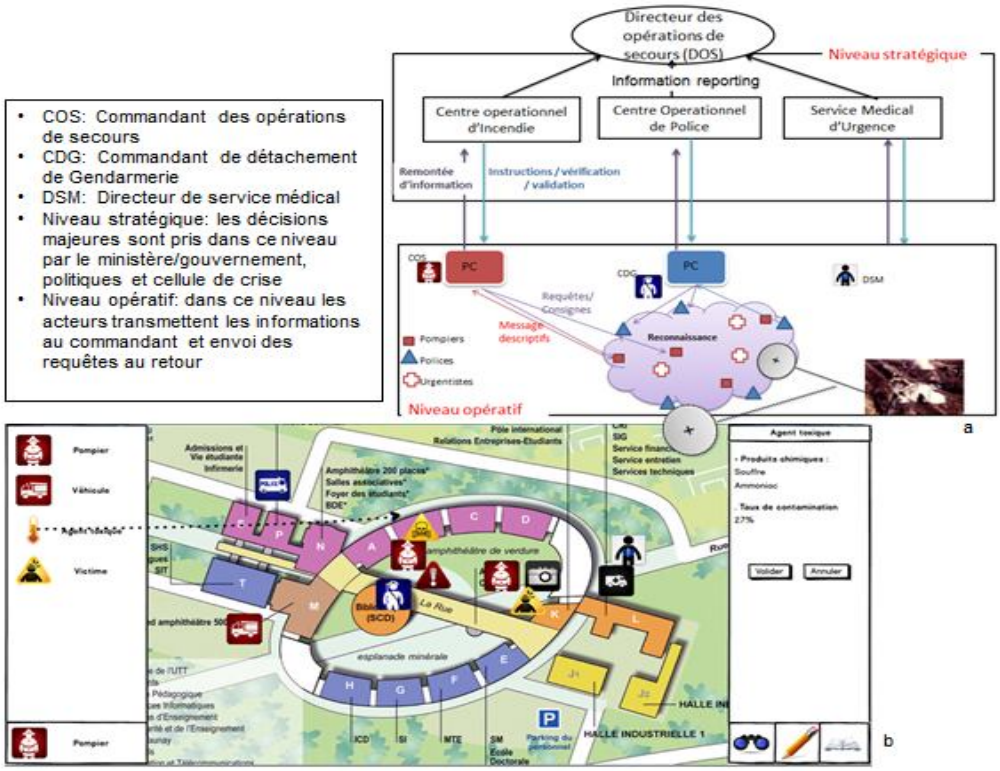


FIGURE 1 – Modèle général de communication et d'intervention

2 Conclusion et perspective

La multitude des organisations intervenantes sur une crise et la différence de leur culture engendre des problèmes de communication transversale inter-organisationnelle sur le terrain, en plus des problèmes d'interprétation de message et d'interopérabilité des systèmes, Les agents de différents service ne communiquent pas suffisamment transversalement pour engendrer une conscience mutuelle ce qui influence la prise de décision. Dans notre analyse, nous avons constaté que la communication transversale des informations utiles pouvait être améliorée dans plusieurs aspects renforçant une bonne gestion de crise. Pour cela nous avons proposé une approche préliminaire d'un système participatif communicatif (qui est en cours), dédiée aux acteurs des services d'urgence. Ce système s'appuie sur un mode participatif pour visualiser les points d'intérêts d'une crise et les localiser sur carte géographique, afin de faciliter la communication d'information et la conscience collective.

Dans la suite de ce travail, nous allons renforcer cette approche et ajouter les interactions des actions des différents acteurs ainsi que la hiérarchisation des données provenant de différentes sources.

Références

Ma X., Cahier J.-P. (2012). Visual Distinctive Language: using a Hypertopic-based Iconic Tagging System for Knolwedge Sharing. IEEE 21st International WETICE Conference, 5th Web2Touch Track (Modeling the Collaborative Web Knowledge Conference), Toulouse (France), June 25-27 2012.

Zhou C., Lejeune Ch. and Bénel A. (2006). Towards a standard protocol for community-driven organizations of knowledge, in Proceedings of the 13th International Conference on Concurrent Engineering (ISPE CE'06), IOS Press.

Saoutal A., Cahier J.-P., Matta N. (2014). Modeling the communication between emergency actors in crisis management. Collaboration Technologies and Systems (CTS), International Conference (À paraître).

Intégration d'un réseau bayésien dans une ontologie

Emna Hlel, Salma Jamoussi et Abdelmajid Ben Hamadou

Laboratoire MIRACL, Pôle Technologique de Sfax BP 242 - 3021, Sakiet Ezzit Sfax, TUNISIE,
emnahlel@gmail.com, jamoussi@gmail.com et abdelmajid.benhamadou@gmail.com

Résumé : L'augmentation et la diversification d'informations ont créé de nouveaux besoins utilisateurs. Les problèmes de représentation, traitement, analyse et raisonnement sur l'informations, et surtout les informations incertaines, constituent encore un thème de recherche important. Donc, il est indispensable de proposer des nouvelles approches permettant de représenter formellement les informations incertaines pour aider les machines à les comprendre et à inférer des nouvelles connaissances. Cet article est inscrit dans ce cadre.

Mots-clés : ontologie probabiliste (OP), ontologie classique, réseau bayésien (RB), inférence, incertain, etc.

1 Introduction

Selon (Studer R. et al., 1998), une ontologie est «*une spécification explicite et formelle d'une conceptualisation partagée*». Elle est une représentation qui regroupe un ensemble de concepts et relations décrivant un domaine particulier. Un de principaux défauts de l'ontologie est leur incapacité de représenter et raisonner sur l'incertitude. Dans les années précédentes, divers chercheurs (Nottelmann H. & Fuhr N., 2006), (da Costa et al.,2005), (Heinsohn J., 1994), (Pool M. & Aikin J., 2004), etc ont tenté de proposer des approches visant à intégrer l'incertitude dans les ontologies, qui est un axe de recherche intéressant sur lequel nous avons situé ce travail. L'une de caractéristiques de RB est leur capacité de représenter les informations incertaines sous la forme d'un modèle probabiliste. Mais, malheureusement cette représentation n'est pas formelle et les machines ne sont pas capables de la comprendre. Pour cette raison, nous tentons de proposer une méthode permettant de représenter formellement un RB sous la forme d'une ontologie probabiliste à l'aide d'un standard de représentation d'ontologie OWL (Web Ontology Language). Cette représentation formelle peut être utilisée par la suite comme un support à des opérations de raisonnement dans des contextes différents comme la recherche de documents sur le Web ou dans l'indexation sémantique des pages Web, etc.

2 Etat de l'art

D'après (Costa et al., 2005), une OP est « *an explicit, formal knowledge representation that expresses knowledge about a domain of application*». Les OPs permettent de décrire les connaissances sur un domaine d'une manière raisonnée, structurée et partageable, idéalement dans un format qui peut être lu et traité par un ordinateur et d'intégrer l'incertitude à ces

connaissances. Un RB est un modèle probabiliste qui est capable de gérer l'incertain. Il est défini par un graphe orienté sans circuit dont les sommets représentent des variables aléatoires d'un domaine, les arcs indiquent des dépendances conditionnelles entre les sommets et des probabilités conditionnelles permettent de quantifier les dépendances entre les nœuds. Il repose sur la théorie de graphe pour représenter les dépendances conditionnelles entre les variables d'un système étudié, et sur la théorie de probabilités pour définir mathématiquement ces dépendances (Philippe L., 2006). Divers chercheurs ont essayé de combiner les RBs avec les ontologies afin de représenter et raisonner avec l'incertitude : (Costa et al., 2005), (Yang Y. & Calmet J., 2005), (Ding & Peng, 2004) ont proposé des extensions de la formalisme standard OWL qui sont BayesOWL, OntoBayes et PR-OWL. (Koller D. & Levy A., 1997), (Fabio G. et al, 2011), etc ont proposé des nouvelles extensions probabilistes de logiques de description probabilistes (LDP) qui sont une famille de langages de représentation de connaissances ontologiques probabilistes et qui s'adoptent bien au Web.

Nous avons intégré le RB dans l'ontologie afin de remédier les défauts de chacune (le RB est incapable de représenter formellement les informations incertaines et l'ontologie est incapable de gérer l'incertitude) et de combiner les inférences bayésiennes (mise à jour de probabilités) avec l'inférence ontologique (vérification d'instances, consistance, etc). L'OP obtenue peut être utilisée par la suite dans des applications d'aide à la décision qui doivent être capable de raisonner avec des connaissances incertaines.

3 Notre méthode de construction d'OP

Un des principaux avantages de RB est leur capacité de gérer l'incertitude : cela permet de représenter les informations incertaines avec une manière très lisible et clair et offre des techniques puissants d'inférences bayésiennes qui consiste à propager une ou plusieurs informations (valeurs probabilistes) dans le réseau pour en déduire comment ceci intervient sur les probabilités d'autres variables. Cette inférence est appelée "mise à jour" des probabilités. Pour formaliser la représentation de ce modèle et la rendre compréhensible par les machines, nous proposons une méthode de modélisation formelle de RB en utilisant la formalisme standard d'ontologie OWL. L'objectif de cette méthode n'est pas seulement de représenter formellement les informations provenant de RB sous la forme d'une OP mais aussi de combiner quelques points forts de RB avec ceux de l'ontologie : les outils de raisonnement sur l'ontologie (subsomption, vérification d'instances, consistance, ...) avec les outils offerts par le RB (représentation des informations incertaines, l'inférence,…) afin de fournir des mécanismes permettant de représenter les informations incertaines et d'inférer des connaissances implicites à partir des connaissances explicites stockées dans une ontologie probabiliste.

Soit Q est un réseau bayésien qui est défini par un graphe G=(A, S), avec A= $\{A_1, A_2, \ldots, A_m\}$ est l'ensemble de relations de dépendance de taille m et S=$\{S_1, S_2, \ldots, S_n\}$ est l'ensemble de sommets de taille n. On observe la similitude de structure entre le graphe de RB et le graphe d'ontologie : Le RB et l'ontologie sont définis par un graphe. Ces deux graphes possèdent un ensemble d'arcs et un ensemble de sommets. Cette similitude facilite la tâche de transformation de RB en une OP qui peut être faite à l'aide d'un ensemble de règles de traduction qui sont :

1) Le n nœuds de graphe de RB sont transformés en un ensemble de concepts C= $\{c_1, c_2, \ldots, c_n\}$, avec n est le nombre de concepts d'ontologie.

2) Les valeurs possibles de chaque nœuds sont transformées en ensemble d'instances de concepts I. I=$\{I_1, I_2, \ldots, I_n\}$ et I_j=$\{i_1, i_2, \ldots, i_k\}$, avec I_j est l'ensemble d'instance de concept j et k est le nombre d'instances de concept j.

3) La probabilité à priori P(A) est converti en une valeur probabiliste (entre 0 et 1) d'une propriété de type DataProperty, nommée «ProbApriori-ci», de concept ci. $ci \in Rac, \text{ avec } Rac = \{c1, c2, \ldots, cl\}$ est l'ensemble de racines de RB (Rac sous ensemble de

C), et l est le nombre de nœuds sans parents (racines). $ci \in Rac$ si et seulement si $ci \in C$ et $Parent(ci) = \emptyset$.

4) La présence d'un arc entre deux nœuds de RB présente une relation de dépendance entre ces deux nœuds. Cette relation de dépendance sera transformée en une relation N-aire (une relation N-aire est une relation ayant plus de deux arguments ou un attribut additionnel.), nommée «dépend-de-j», qui est caractérisée par une propriété de type DataProperty , nommée «ProbCond-j». Cette propriété exprime la probabilité conditionnelle P(A|B), qui indique la probabilité pour chaque valeur du A, étant donné les combinaisons de valeurs de parents du A (B). Donc, le m arcs seront convertis en un ensemble de relations N-aire entre les concepts $R=\{R_1$: dépend-de-1, R_2 : dépend-de-2, …, R_m : dépend-de-m$\}$. Chaque relation R_j de R est caractérisée par une propriété de type DataProperty : «ProbCond-j».

4 Conclusion

Dans cet article, nous avons proposé une méthode permettant de représenter sémantiquement un RB sous la forme d'une OP. Dans le futur travail, nous essayerons d'utiliser cette OP dans des applications réelles et nous nous intéresserons plus à l'inférence dans l'OP : Quels sont les axiomes (ajoutés dans OP) que nous pouvons déduire à partir de RB ? Quelles sont les nouvelles connaissances que nous pouvons extraire ou inférer à partir de cette ontologie probabiliste ?...

Références

Studer R. et al. (1998). Studer R. Benjamins, V.R., Fensel, D., Knowledge engineering: Principles and methods. Data & Knowledge Engineering 25, 161–197.

Nottelmann H. & Fuhr N. (2006). Adding probabilities and rules to OWL Lite subsets based on probabilistic Datalog, Int. J. Uncertain. Fuzz. pp 17–42.

Heinsohn J. (1994). Probabilistic description logics, in: Proceedings UAI-1994, Morgan Kaufmann, pp. 311–318.

Pool M. & Aikin J. (2004). KEEPER and Protégé: An elicitation environment for Bayesian inference tools, in: Proceedings of the Workshop on Protégé and Reasoning held at the IPC.

da Costa P.C.G. et al. (2005). da Costa P.C.G. et Laskey K.B. Laskey K.J., PR-OWL: A Bayesian ontology language for the Semantic Web, in: Proceedings URSW, pp. 23–33.

Philippe L. (2006). Réseaux bayésiens : apprentissage et modélisation de systèmes complexes, HABILITATION.

Yang Y. & Calmet J. (2005). Ontobayes: An ontology-driven uncertainty model. In: (IAWTIC'05). IEEE Computer Society (2005) 457-464.

Ding Z. & Peng Y. (2004). A probabilistic extension to ontology language OWL, in: Proceedings HICSS.

Koller D. & Levy A. (1997). P-CLASSIC: A tractable probabilistic description logic, in: Proceedings AAAI-1997, AAAI Press/MIT Press, pp. 390–397.

Fabio G. et al (2011). Rodrigo B. Polastroa, Fabio G. Cozmana, Felipe I. Takiyamaa, and Kate C. Revoredob, Computing Inferences for Credal ALC Terminologies.

Stanjanovic L. (2004). Methods and Tools for Ontology Evolution, Thèse de doctorat de l'Université de Karlsruhe.

Index des auteurs

Marie-Hélène Abel 263
Céline Alec 87
Pavel Arapov 249, 283
Ala Atrash 263
Bruno Bachimont 63
Abdelmajid Ben Hamadou 295
Amel Ben Othmane 249, 283
Sadok Ben Yahia 187
Nadia Bennani 287
Uriel Berdugo 87
Brigitte Blaszka-Jaulerry 107
Alain Bonardi 63
Jacques Bouaud 107
Patrick Brébion 267
Sandra Bringay 163
Michel Buffa 249, 283
Jean-Pierre Cahier 267, 291
Gaoussou Camara 39
Stefano A. Cerri 175
Pierre-Antoine Champin 137
Jean Charlet 15, 75
Michel Chein 225
Amina Chniti 75
Isabelle Cojean-Zelek 107
Olivier Corby 213, 237
Olivier Coupelon 271
Christopher Couthon 119
Michel Crampes 151
Madalina Croitoru 225
Christel Daniel 75
Stefan Darmoni 75
Gunnar Declerck 15
Sylvie Despres 27, 39
Diyé Dia 271
Catherine Duclos 51
Axel Durieux 107
Elöd Egyed-Zsigmond 287
Benoît Encelle 125
Gilles Falquet 99
Catherine Faron-Zucker 213, 237
Alban Gaignard 237
Fabien Gandon 279
Emmanuel Gaude 287
Pierre Gayet 15
Alain Giboin 279
Blandine Ginon 137
Nicolas Griffon 75
Romain Gugert 279
Léa Guizol 275
Ollivier Haemmerlé 201
Nathalie Hernandez 201
Emna Hlel 295
Sajjad Hussain 75
Stéphanie Jean-Daubias 137
Salma Jamoussi 295
Mohamed Nader Jelassi 187
Clément Jonquet 175
Abir Beatrice Karami 125
Eric Kergosien 163
Philippe Lamarre 287
Jean-Baptiste Lamy 51
Michel Leclère 225
Marie Lefevre 137
Jean-Pierre Lefranc 107
Maxime Lefrançois 279
Mylène Leitzelman 267
Philippe Lemoisson 175
Eric Lepage 75
Moussa Lo 39
Yannick Loiseau 271
Abdelkarim Mars 299
Régis Martineau 119
Nada Matta 291
Laurent Mazuel 15
Nizar Messai 107
Patrick Miroux 15
Engelbert Mephu Nguifo 187
Claude Moulin 263
David Ouagne 75
Nathalie Pernelle 225
Michel Plantié 151
Pierre Pompidor 163

Pascal Poncelet …….. 163
Camille Pradel ……... 201
Olivier Raynaud …… 271
Chantal Reynaud ……. 87
Hélène de Ribaupierre ……. 99
Alexandra Rousseau …….. 107
Eric Sadou ……... 75
Brigitte Safar ……... 87
Fatiha Saïs ……. 225
Pascal Salembier …… 119
Amina Saoutal ……... 291
Karim Sehaba …… 125
Zied Sellami ……. 87
Brigitte Seroussi …… 107
Oumy Seye …… 237
Lina F. Soualmia ……. 51
Arnaud Soulet …… 107
Jean-Philippe Spano …….. 107
Guillaume Surroca ……. 175
Danai Symeonidou …… 225
Christophe Tournigand …….. 107
Lamine Traore ……..… 75
Alain Venot ……. 51
Antoine Vincent …….. 63
Laurent Zelek …… 107