Actes des 28es journées francophones d'Ingénierie des Connaissances

Catherine Roussey

Actes IC 2017

28es Journées francophones d'Ingénierie des Connaissances

PFIA17

plate-forme d'intelligence artificielle

3 au 7 juillet 2017, Caen

Comité de programme de IC 2017

Présidente

Catherine ROUSSEY, TSCF, Irstea Clermont Ferrand, France.

Relecteurs :

- Marie-Hélène ABEL, HEUDIASYC, Université de Technologie de Compiègne, France.
- Xavier AIME, Cogsonomy Nantes/ LIMICS Paris, France.
- Yamine AIT AMEUR, IRIT, INP Toulouse, France.
- Florence AMARDEILH, Mondeca, Paris, France.
- Fabien AMARGER, IRIT, Toulouse, France.
- Nathalie AUSSENAC-GILLES, IRIT, CNRS, Toulouse, France.
- Bruno BACHIMONT, COSTECH, Université de Technologie de Compiègne, France.
- Nacéra BENNACER, LRI, Centrale Supélec campus de Gif-sur-Yvette, France.
- Aurélien BENEL, ICD, Université de Technologie de Troyes, France.
- Pierre BOURHIS, LIFL, CNRS, Lille, France.
- Nathalie BRICON-SOUF, IRIT, Université Paul Sabatier Toulouse 3, France.
- Sandra BRINGAY, LIRMM, Université Paul-Valéry Montpellier 3, France.
- Patrice BUCHE, IATE, INRA Montpellier, France.
- Davide BUSCALDI, LIPN, Université Paris 13, France.
- Elena CABRIO, I3S, Université de Nice Sophia Antipolis, France.
- Sylvie CALABRETTO, LIRIS, INSA de Lyon, France.
- Gaoussou CAMARA, Université Alioune Diop de Bambey, Sénégal.
- Pierre-Antoine CHAMPIN, LIRIS, Université Claude Bernard Lyon 1, France.
- Jean-Pierre CHANET, TSCF, Irstea Clermont-Ferrand, France.
- Jean CHARLET, LIMICS, AP-HP/INSERM, Paris, France.
- Olivier CORBY, I3S, INRIA Sophia Antipolis-Méditerranée, France.
- Amélie CORDIER, Hoomano, Lyon, France.
- Mathieu D'AQUIN, Knowledge Media Institute, Open University, RU.
- Luc DAMAS, LISTIC, Université de Savoie, France.
- Jérôme DAVID, LIG,INRIA Grenoble, France.
- Sylvie DESPRES, LIMICS, Université Paris 13, France.
- Rim DJEDIDI, LIMICS, Université Paris 13, France.
- Jean-Pierre EVAIN, EBU, Suisse.
- Gilles FALQUET, Université de Genève, Suisse.
- Catherine FARON-ZUCKER, I3S, Université de Nice Sophia Antipolis, France.
- Cécile FAVRE, ERIC, Université Lumière Lyon 2, France.
- Béatrice FUCHS, LIRIS, Université Jean Moulin Lyon 3, France.
- Frédéric FURST, MIS, Université de Picardie Jules Verne, France.
- Fabien GANDON, I3S, INRIA Sophia Antipolis Méditerranée, France.
- Jean-Gabriel GANASCIA, LIP6, Université Pierre et Marie Curie, France.
- Catherine GARBAY, LIG, CNRS, Grenoble, France.
- Serge GARLATTI, IMT Atlantique, Brest, France.
- Alain GIBOIN, I3S, INRIA Sophia Antipolis - Méditerranée, France.
- Nathalie GUIN, LIRIS, Université Claude Bernard Lyon 1, France.
- Ollivier HAEMMERLE, IRIT, Université Toulouse le Mirail, France.
- Mounira HARZALLAH, LS2N, Université de Nantes, France.
- Nathalie HERNANDEZ, IRIT, Université Toulouse Le Mirail, France.

- Liliana IBANESCU, MIA, INRA AgroParistech, Paris, France.
- Antoine ISAAC, Europeana & VU University Amsterdam, Pays-Bas.
- Clément JONQUET, LIRMM, Université de Montpellier, France.
- Mouna KAMEL, IRIT, Université de Perpignan Via Domitia, France.
- Gilles KASSEL, MIS, Université de Picardie Jules Verne, France.
- Pascale KUNTZ, LS2N, Université de Nantes, France.
- Florence LE BER, ICUBE, ENGEES, Strasbourg, France.
- Michel LECLERE, LIRMM, Université de Montpellier, France.
- Maxime LEFRANCOIS, Lab. Hubert Curient, Ecole des Mines de Saint-Etienne, France.
- Alain LEGER, Orange Labs, France Telecom, Rennes, France.
- Dominique LENNE, HEUDIASYC, Université de Technologie de Compiègne, France.
- Moussa LO, Université Gaston Berger de Saint Louis, Sénégal.
- Cédric LOPEZ, Viséo Objet Direct, Grenoble, France.
- Nada MATTA, ICD, Université de Technologie de Troyes, France.
- Pascal MOLLI, LS2N, Université de Nantes, France.
- Alexandre MONNIN, I3S, INRIA Sophia Antipolis Méditerranée, France.
- Fleur MOUGIN, INSERM BPH U1219, Université de Bordeaux, France.
- Amedeo NAPOLI, LORIA, CNRS, Nancy, France.
- Emmanuel NAUER, LORIA, Université de Lorraine, France.
- Jérôme NOBECOURT, LIMICS, Université Paris 13, France.
- Nathalie PERNELLE, LRI, Université Paris Sud, France.
- Camille PRADEL, Synapse, Toulouse, France.
- Yannick PRIÉ, LS2N, Université de Nantes, France.
- Cédric PRUSKI, Luxembourg Institute of Science and Technology, Luxembourg.
- Sylvie RANWEZ, LGI2P, Ecole des mines d'Alès, France.
- Chantal REYNAUD, LRI, Université Paris Sud, France.
- Catherine ROUSSEY, TSCF, Irstea Clermont-Ferrand, France.
- Fatiha SAIS, LRI, Université Paris Sud, France.
- Pascal SALEMBIER, ICD, Université de Technologie de Troyes, France.
- Karim SEHABA, LIRIS, Université Lumière Lyon 2, France.
- Hassina SERIDI-BOUCHELAGHEM, LabGED, Université de Badji Mokhtar, Algérie.
- Andrea TETTAMANZI, I3S, Université Nice Sophia Antipolis, France.
- Raphaël TRONCY, EURECOM, Sophia Antipolis, France.
- Serena VILLATA, I3S, CNRS, France.
- Amel YESSAD, LIP6, Université Pierre et Marie Curie, Paris, France.
- Haïfa ZARGAYOUNA, LIPN, Université Paris 13, France.
- Pierre ZWEIGENBAUM, LIMSI, CNRS, Université Paris-Saclay, Orsay, France.

Relecteurs additionnels:

- Sahar ALJALBOUT, Université de Genève, Suisse.
- Nicolas SEYDOUX, LAAS et IRIT, Toulouse, France
- Elodie THIEBLIN, IRIT, Toulouse, France.

Préface

La sélection d'articles publiés dans ce recueil constitue les actes des 28es Journées Francophones d'Ingénierie des Connaissances (IC 2017). Cette conférence s'est déroulée du 3 au 7 juillet 2017 à Caen, dans le cadre de la Plate-Forme Intelligence Artificielle (PFIA) : https://pfia2017.greyc.fr/rjcia.

Organisées chaque année depuis 1997 sous l'égide du Gracq (Groupe de Recherche en Acquisition des Connaissances) puis du collège IC de l'AFIA, les journées francophones d'Ingénierie des Connaissances constituent un lieu d'échanges et de réflexions de la communauté francophone. Chercheurs académiques, industriels et étudiants s'y retrouvent pour échanger sur les concepts, méthodes et techniques permettant de modéliser, d'acquérir et de traiter les connaissances dans des domaines d'application variés.

Ces journées avaient pour thème "*Web des connaissances ou évolution des systèmes à base de connaissances face aux avancées des technologies Web (Web des Objets, Web des Données, Service Web)*". En effet, les technologies Web ont considérablement modifié les pratiques de conception de systèmes d'information. Les organisations publiques ou privées publient sur le Web leurs données de référence et les jeux de données sont maintenant liés les uns aux autres. Les entreprises exposent leurs services de traitement sous forme d'API Web pour faciliter leur réutilisation par d'autres systèmes. Les objets physiques ont aussi leur pendant virtuel sur le Web pour être intégrés dans des systèmes adaptatifs et contextuels. La question de la place des connaissances, de leur gestion et des modèles associés est centrale au sein de ces systèmes qui combinent une multitude de données, de traitements et d'objets liés les uns aux autres. En quoi les connaissances peuvent améliorer le développement de systèmes d'information de plus en plus complexes ? Partager des connaissances au sein d'une communauté, d'une entreprise ou d'un système d'information suppose leur explicitation, leur représentation, leur exploitation, leur mise en relation, leur diffusion, leur maintenance. L'ingénierie des connaissances est au cœur de ces problématiques.

Pour illustrer ce thème Web des connaissances, nous avons reçu comme conférencier invité Raùl Garcia-Castro du laboratoire Ontology Engineering Group de Madrid. Sa présentation portait sur l'interopérabilité sémantique pour l'Internet des objets.

30 soumissions d'articles ont été reçues. Chaque soumission d'article a été relue par au moins 3 relecteurs. Le comité a décidé d'accepté 14 papiers longs (taux d'acceptation de 47%) et 10 papiers courts (taux d'acceptation totale de 80 %). Les 4 démonstrations reçues ont toutes été admises. Les démonstrations ont été présentées lors d'une session commune de la plateforme PFIA.

Les articles de cette édition ont été regroupés en plusieurs sessions thématiques :

- Ingénierie des connaissances médicales
- Alignements d'ontologies pour le Web de données
- Extraction de connaissances
- Développement d'ontologies
- Graphes et connaissances
- Qualité, vérité et recommandations.

Chacune de ces sessions constitue un chapitre de ce recueil. Le chapitre contenant les articles associés aux démonstrations conclut les actes de IC 2017.

Je remercie vivement les auteurs pour leurs contributions, le comité de programme et les relecteurs additionnels pour la qualité de leurs recommandations ainsi que les organisateurs de la Plate-Forme d'Intelligence Artificielle pour leur investissement. Mon travail de présidente de programme a été largement facilité par l'utilisation du système web easy-chair.

Catherine Roussey, *Présidente du comité de programme*.

Sommaire

Développement d'ontologies

Graphes et connaissances

Qualité, vérité et recommandation

Démonstrations

Index des auteurs

Technical and social aspects of semantic interoperability in the IoT

Raúl García-Castro 1

[1] ONTOLOGY ENGINEERING GROUP, Universidad Politécnica de Madrid, Madrid, Spain
rgacia@fi.upm.es

Abstract

The Internet of Things (IoT) envisions an ecosystem in which physical entities, systems and information resources bridge the gap between the physical and the virtual world. The existing heterogeneity in such physical entities, systems and information resources, intensified by the fact that they originate from different sectors and according to different perspectives, poses numerous challenges to the IoT vision.

One of them is the need for interoperability, since capturing the maximum value from the IoT involves multiple IoT systems working together and, therefore, seamlessly interchanging information. However, successfully achieving interoperability requires coping with different aspects, not only technological but also social and/or regulatory ones. This talk will address how these aspects influence semantic interoperability, taking into account that such interoperability requires being aware of both the information interchanged and the data model (i.e., ontology) of such information.

In order to achieve interoperability, systems need not only to successfully interchange information but also to use the information that has been interchanged. In the IoT, semantic interoperability not only requires interchanging the information itself, but also the ontologies used to represent such information and other types of information that support IoT-specific tasks (e.g., discovery). Furthermore, using the interchanged information will require, on the one hand, to understand the information (usually through an ontology) and, on the other hand, to deal with mismatches among different views of the world.

The latter is very important because the reality is that the landscape of IoT ontologies is fragmented and reconciling views goes beyond solving technical issues and requires social approaches. This need in the IoT field for consensual models has led to multiple initiatives to define consensus-driven ontologies both in standardisation bodies and in other groups that aim to produce de facto standards. Even so, these processes require a special focus on aspects such as collaborativeness or openness that are partially tackled with in traditional ontological engineering practices and tools and bring new demands for them.

This talk will discuss current approaches and challenges for semantic interoperability in the IoT, covering not only technical aspects but also social ones, presented through different examples drawn from the VICINITY H2020 project and various initiatives in ontology standardisation.

Formalisation de la terminologie LOINC® et évaluation de ses avantages pour la classification des tests de laboratoire

Mélissa Mary[1,2], Lina F. Soualmia[2,3] et Xavier Gansel[1]

[1] bioMérieux SA, Dépt. Développement et Intégration, 38390 La Balme Les Grottes,
{melissa.mary ;xavier.gansel}@biomerieux.com
[2] Normandie Universités, UNIVROUEN - LITIS EA 4108, 76000 Rouen
Lina.Soualmia@chu-rouen.fr
[3] LIMICS INSERM UMR_1142, 75000 Paris,

Résumé : La numérisation croissante des données médicales a fait émerger de nouveaux standards pour représenter l'information au sein des dossiers patients informatisés. Dans le domaine du diagnostic *in vitro*, deux standards sémantiques sont recommandés pour coder les descriptions des tests de laboratoire (la terminologie LOINC®) et les résultats obtenus (l'ontologie SNOMED CT®). La structuration non formelle des tests et l'absence de relations hiérarchiques au sein de LOINC® limitent les capacités d'interrogation et de réutilisation des données du diagnostic *in vitro* au sein des dossiers patients. L'objectif de ce travail est d'évaluer les avantages d'une représentation formelle permettant de décrire et de classer automatiquement les tests de laboratoire. Dans cet article, nous présentons ontoLOINC, une représentation ontologique de LOINC® et les processus de génération que nous avons mis en place. Dans un second temps, nous comparons les classifications des tests dans ontoLOINC et dans un formalisme fondé sur SNOMED CT® (LOINC—SNOMED CT). La formalisation des tests et la classification des éléments utilisés dans leur définition permet d'obtenir des classifications cohérentes avec la terminologie.

Mots-clés : modélisation d'ontologie – peuplement d'ontologie – classification automatique – évaluation, LOINC® – SNOMED CT®

1 Introduction

L'informatisation croissante des données médicales au sein de dossiers médicaux partagés a pour objectif d'améliorer la prise en charge du patient par l'établissement d'échanges d'informations interopérables entre acteurs et systèmes de la chaîne de soins (Macary, 2007; Stroetmann, 2009). Une des clés de l'interopérabilité repose sur l'utilisation des vocabulaires standards pour coder l'information au sein des dossiers patients. Ces vocabulaires spécialisés dans un domaine peuvent être représentés sous forme de terminologie, thesaurus, ou ontologie. Dans le domaine du diagnostic *in vitro*[1] deux ressources sont recommandées par les instances de standardisation nationales et internationales pour coder les informations au sein des comptes-rendus d'analyses (Blumenthal, 2010; Stroetmann, 2009). La terminologie LOINC®[2] (*Logical Observation Identifiers Names and Codes*) est préconisée pour décrire les tests de laboratoire, et l'ontologie SNOMED CT®[3] (*Systematized Nomenclature of MEDicine – Clinical Terms*) pour coder les résultats obtenus. Les informations exprimées par LOINC® et SNOMED CT® dans les dossiers patients sont interdépendantes : l'interprétation d'un résultat varie en fonction du test qui a permis de l'obtenir. LOINC® et SNOMED CT®

[1] Le diagnostic *in vitro* regroupe l'ensemble des analyses biologiques réalisées sur un échantillon clinique (sang, urine) qui permettent de détecter et caractériser une pathologie
[2] http://loinc.org/
[3] http://www.snomed.org/

doivent donc être utilisées conjointement pour permettre l'interrogation et l'agrégation des données issues d'un compte-rendu d'analyse. L'absence de représentation formelle et l'absence de relations hiérarchiques entre les tests LOINC® limitent les capacités d'interrogation et de réutilisation des données du diagnostic *in vitro* au sein des dossiers patients. Pour répondre à cette problématique, le *Regenstrief Institute* et l'IHTSDO (maintenant appelé SNOMED International) ont mis en place une collaboration en 2013 (IHTSDO & Regenstrief Institute, 2013; Vreeman, 2015) qui a pour vocation de rendre interopérable les données du compte-rendu de laboratoire en proposant une ressource alignant les tests LOINC® avec l'ontologie SNOMED CT® (ressource LOINC—SCT).

L'objectif de ce travail est d'étudier l'impact d'une formalisation logique de la représentation des tests de diagnostic *in vitro,* initialement décrit par la terminologie LOINC. Dans un premier temps, nous avons développé ontoLOINC une ontologie représentant la terminologie LOINC® ; elle pour pallier aux contraintes inhérentes à la ressource LOINC—SCT. Dans une seconde partie nous nous comparons ontoLOINC et LOINC—SCT en étudiant les inférences obtenues à partir des deux modèles. Cette étude permet notamment d'illustrer les bénéfices d'une formalisation pour la classification des LOINC®.

Cet article est organisé comme suit : dans la section 2 nous décrivons les ressources LOINC®, SNOMED CT® et l'alignement LOINC—SCT qui ont été utilisés pour cette étude. La section 3 présente ontoLOINC, son processus de génération et le protocole de classification automatique des tests. La section 4 présente les résultats de classification automatique sur les représentations ontoLOINC et LOINC—SCT. Nous discutons les résultats obtenus en section 5 et concluons ce travail en section 6.

2 Ressources utilisées

2.1 Logical Observation Identifiers Names and Codes

La terminologie LOINC® a été construite en 1994 pour décrire les tests cliniques ou réalisés par le laboratoire d'analyse sur un prélèvement (sang, urine, *etc.*). Elle est mise à jour deux fois par an par le Regenstrief Institute et est couverte par une licence d'utilisation gratuite. La terminologie est accessible en ligne ou via le logiciel RELMA. (McDonald *et al.*, 2003; Sheide & Wilson, 2013).

Dans cette étude nous utilisons la version 2.52 de la terminologie (décembre 2015) ; elle décrit 70 789 tests actifs dans une structure normée par 6 dimensions obligatoires et 3 optionnelles. La description d'un test représente à la fois des informations sur le protocole expérimental (*Composant*, *Milieu*, *Technique* et *Temps*) et sur le type de résultat attendu (*Échelle* et *Unité*). Dans cet article nous utilisons le terme générique *partie* pour parler de la valeur d'une dimension utilisée pour décrire un test.

Table 1 Dimensions du modèle LOINC®

Dimension	Description	Exemple
Composant	Objet du test ou caractéristique testée	*Bacteria identified*
Milieu	Echantillon sur lequel est réalisé le test	*Bld* (*Blood)*
Méthode	Technique d'analyse	*IF (Immunofluorescence)*
Temps	Temporalité du test (cinétique ou ponctuelle)	*Pt* (*ponctuel)*
Échelle	Type de résultats	*Qn pour quantitatif*
Unité	Description précise de l'unité du résultat	*MCnc* (*Mass/volume)*
Challenge	Contexte de la réalisation du test	
Diviseur	Composant utilisé dans le cas où le résultat du test correspond un ratio	
Ajustement	Paramètre d'ajustement du résultat de la mesure.	

Un test est également contextualisé par des informations complémentaires comme la *classe* qui spécifie un sous type de test (MICRO pour microbiologie, SERO pour la serologie, *etc.*).

2.2 Systematized Nomenclature Of MEDicine Clinical Terms

SNOMED CT® est une ontologie sous licence créée en 2002 pour décrire l'ensemble du domaine clinique (Bhattacharyya, 2016; Cornet & de Keizer, 2008). Elle est maintenue à jour bi-annuellement par SNOMED International (anciennement appelé IHTSDO) et est diffusée dans un système de fichiers plats nommé RF2. SNOMED International fournit également un script permettant de transformer les données du format RF2 en ontologie OWL EL (Bodenreider *et al.*, 2007; Schulz *et al.*, 2009).

L'ontologie est composée de 350 000 concepts répartis en 19 axes. Les axes permettent d'organiser l'ontologie en sous-domaine de connaissance. Par exemple, l'axe *123037004*| *Body Structure*| regroupe l'ensemble des concepts relatifs à la description anatomique et cellulaire du corps humain. Les concepts sont organisés en poly-hiérarchies par des relations de subsomption. Au sein de l'ontologie, on distingue les concepts :

1 pré-coordonnées : ils sont gérés par SNOMED International et constituent le corps de la ressource ;
2 post-coordonnées : ils sont définis pour une utilisation particulière à partir de concepts pré-coordonnées et n'appartiennent pas à la ressource SNOMED CT (SNOMED International, 2017a).

Le concept *Left arm fracture* n'existe pas nativement dans SNOMED CT® ; il peut être créée par post-coordination (Figure 1) à partir des concepts (*Left upper arm structure, Fracture)* et des relations *(Finding site*, *Associated morphology)* existants dans la ressource.

Class: '*Left Arm Fracture'*

EquivalentTo *sct:'Disorder of Bone'* and
sct:associated Morphology some *sct:Fracture* and
sct:finding Site some *sct:'Left upper arm structure'*

Figure 1 Post-coordination du concept *Left arm fracture* en syntaxe Manchester

2.3 Description des tests de laboratoire en SNOMED CT®

Dans cette étude nous utilisons une ressource (LOINC—SCT) qui associe dans un même format LOINC® et SNOMED CT®. Cette dernière est créée par des experts des deux standards et diffusée depuis avril 2016[4] aux formats RF2 et OWL (IHTSDO & Regenstrief Institute, 2013). Cette ressource se compose de deux éléments. Le premier élément décrit l'alignement entre les parties[5] LOINC® et concepts SNOMED CT®.

Cet alignement est utilisé par la suite pour définir formellement les tests LOINC® en concepts post-coordonnés SNOMED CT® (second élément de la ressource). Les tests LOINC® sont représentés par des concepts définis ; ils utilisent les patrons de conceptions développés par un groupe de travail de SNOMED International (figure 2, SNOMED International, 2017b).

La troisième version[4] est une version préliminaire qui couvre 13 756 tests LOINC, soit 20% de la terminologie. La version finale proposera une description de l'ensemble des tests LOINC® utilisés pour représenter les tests de laboratoire et les tests cliniques *Vital Sign*.

[4]http://loinc.org/news/alpha-phase-3-edition-of-draft-loinc-snomed-ct-mappings-and-expression-associations-now-available.html/

[5] valeurs des dimensions d'un test

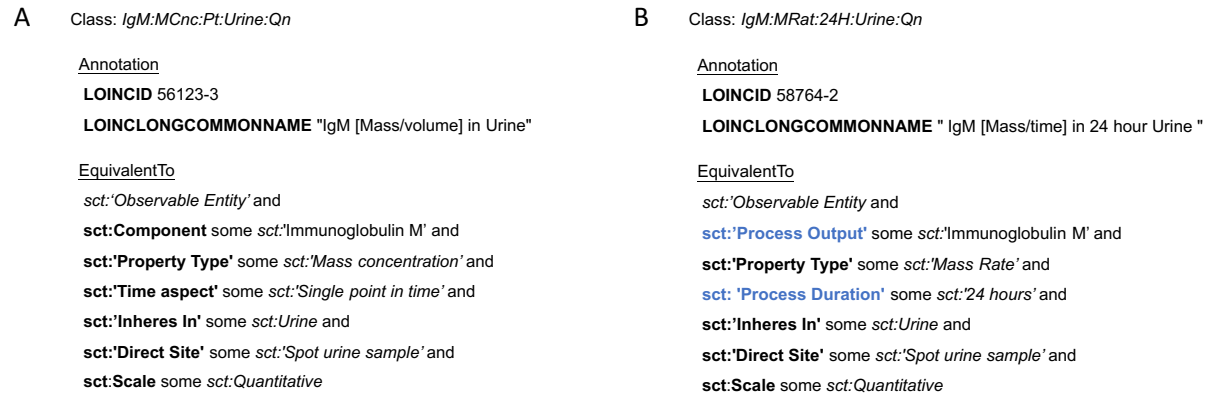


Figure 2 Exemple de formalisation d'un test issue de la terminologie LOINC® avec le patron de conception général (A) et le patron de tests processus (B). Les attributs spécifiques au patron processus sont colorés en bleu

3 Méthodes

3.1 ontoLOINC et processus de génération

L'utilisation de la formalisation des tests LOINC® avec SNOMED CT® est limitée. En effet, seul 20% des tests LOINC® sont représentés dans LOINC—SCT. D'autre part, seules les institutions qui possèdent une licence d'utilisation de l'ontologie SNOMED CT® peuvent utiliser cette ressource pour développer de nouvelles applications. C'est pourquoi nous avons créé une nouvelle ressource, ontoLOINC, que nous décrivons dans cette section.

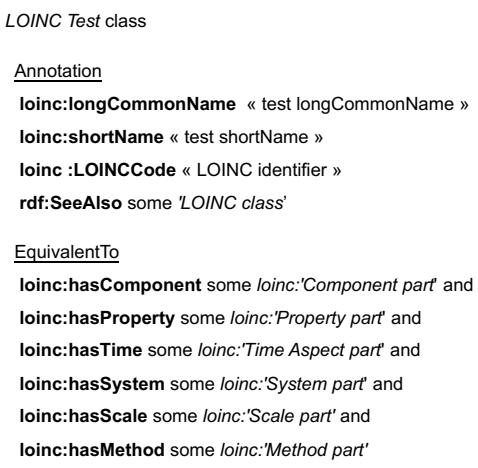


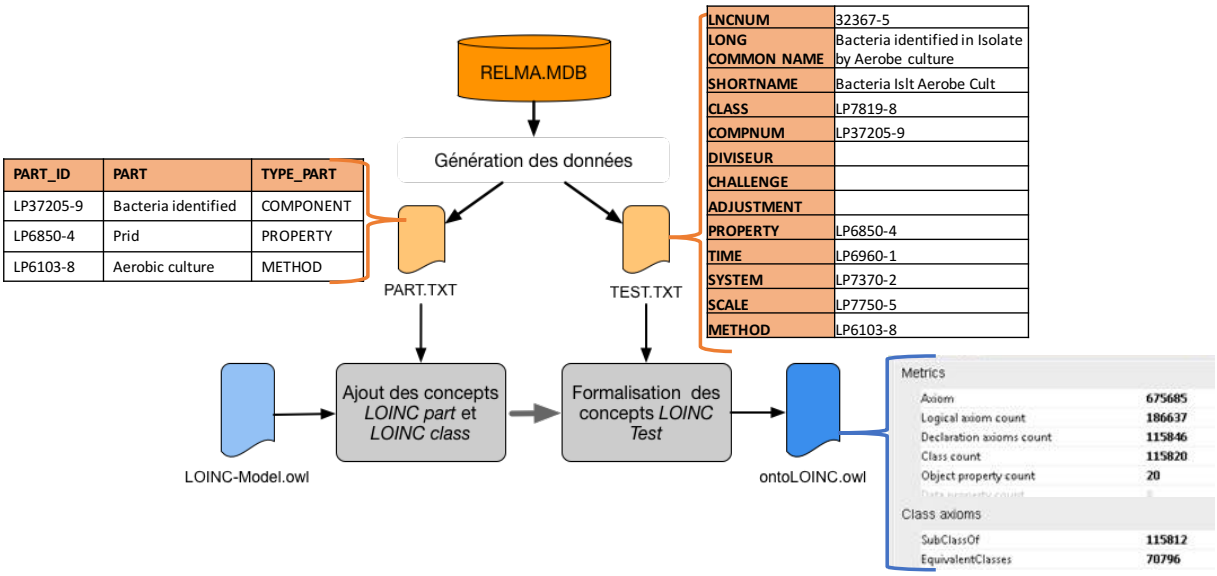
Figure 3 Patron de formalisation d'un test dans ontoLOINC (LOINC Modèle)

Les concepts sont organisés dans ontoLOINC au sein de trois hiérarchies. *loinc:'LOINC part'* contient l'ensemble des concepts représentant des parties ; ils sont classés par dimension. *loinc:'LOINC class'* contient l'ensemble des concepts permettant de contextualiser un test par rapport à son utilisation (classe du test). La hiérarchie *loinc:'LOINC test'* inclut l'ensemble des concepts représentant les tests LOINC. Dans ontoLOINC, nous décrivons un test LOINC® (Figure 3) avec des informations terminologiques (**loinc:longCommonName**, **loinc:shortName**) et une définition formelle. La structuration d'un test LOINC® spécifie la dimension du test à la fois par la relation utilisée (en distinguant une relation par dimension) mais aussi l'espace d'arrivée des relations dans le patron (co-domaine, ou range).

Nous avons développé le processus de génération d'ontoLOINC en trois étapes (Figure 4). Tout d'abord, les *parties* et *tests* sont extraits de la base RELMA (version 2.52) à l'aide de requêtes SQL et stockés dans les fichiers PART.TXT et TEST.TXT. L'algorithme de peuplement d'ontoLOINC est implémenté à l'aide d'un programme java en deux étapes. Dans un premier temps, ontoLOINC est peuplé par l'ensemble des concepts *partie* et *classe*. L'URI de ces concepts est construite à partir de l'identifiant stocké dans le champs PART_ID du fichier PART.TXT. Les concepts sont répartis dans les hiérarchies *loinc:'LOINC class' loinc:'LOINC part'* grâce aux informations stockée dans le champs TYPE_PART. Les tests sont ensuite formalisés sous forme de

concepts définis à partie de des informations décrite dans le fichier TEST.TXT ainsi les concepts *partie* et *classe*.

L'ontologie ainsi créée représente les 70 795 tests actifs de la version 2.52 de la terminologie et 45 004 concepts parties et classes.



PART_ID	PART	TYPE_PART
LP37205-9	Bacteria identified	COMPONENT
LP6850-4	Prid	PROPERTY
LP6103-8	Aerobic culture	METHOD

LNCNUM	32367-5
LONG COMMON NAME	Bacteria identified in Isolate by Aerobe culture
SHORTNAME	Bacteria Islt Aerobe Cult
CLASS	LP7819-8
COMPNUM	LP37205-9
DIVISEUR	
CHALLENGE	
ADJUSTMENT	
PROPERTY	LP6850-4
TIME	LP6960-1
SYSTEM	LP7370-2
SCALE	LP7750-5
METHOD	LP6103-8

Figure 4 Processus de création d'ontoLOINC

3.2 Inférences sur ontoLOINC et LOINC-SCT

3.2.1 Construction de l'ontologie LOINC—SCT

L'ontologie LOINC—SCT est construite en deux étapes. Dans un premier temps, le système de fichier RF2 de la SNOMED CT (janvier 2016) est transformé en ontologie grâce au script perl distribué par SNOMED International. La ressource LOINC—SCT au format OWL est ensuite fusionnée avec l'ontologie SNOMED CT®.

3.2.2 Protocole d'inférence et analyse des données

L'inférence a été exécutée avec le raisonneur ELK (Kazakov *et al.*, 2012) ; il a été choisi pour sa capacité à raisonner sur des ontologies OWL EL volumineuses. Il nous a permis d'obtenir la liste des tests équivalents (4.1) et la classification automatique des tests.

Durant l'analyse des réorganisations hiérarchiques des tests LOINC, nous pouvons distinguer deux types de reclassifications (Figure 5): la surDéfinition et la sousRestriction (Mary *et al.*, 2017; Mary *et al.*, 2016). La **surDéfinition** spécifie toute classification résultant de l'ajout d'une restriction[6] dans la définition formelle du test subsumé. Dans l'exemple ci-dessous le test *20789-4 |Escherichia coli O157 identified in Isolate by Organism specific culture|* (B) est spécifié par une restriction portant sur la dimension *Méthode* non présente chez le test parent *20789-4| Escherichia coli serotype [Identifier] in Isolate* (A). La **sousRestriction** représente toute classification résultant d'une spécialisation d'au moins une restriction de la définition du test subsumé. La spécialisation d'une restriction est portée par le concept utilisé ; celui-ci est subsumé de manière directe ou indirecte par le concept utilisé dans la restriction du test parent (Figure 5 C).

[6] une restriction correspond à une partie de l'axiome définissant un concept test.

L'analyse des classifications est réalisée automatiquement avec un script R. Celui-ci compare entre subsumeur et subsumé la valeur des parties de chaque dimension.

A Class loinc:*'Escherichia coli serotype [Identifier] in Isolate'*

EquivalentTo *loinc:'LOINC test'* and
loinc:hasComponent some loinc:*'Escherichia coli serotype'* and
loinc:hasProperty some loinc:*Prid* and
loinc:hasTime some loinc:*Pt* and
loinc:hasSystem some loinc:*Isolate* and
loinc:hasScale some loinc:*Nom*

B Class *loinc:'Escherichia coli O157 identified in Isolate by Organism specific culture'*

EquivalentTo *loinc:'LOINC test'* and
loinc:hasComponent some loinc:*'Escherichia coli O157 identified '* and
loinc:hasProperty some loinc:*Prid* and
loinc:hasTime some loinc:*Pt* and
loinc:hasSystem some loinc:*Isolate* and
loinc:hasScale some loinc:*Nom*
loinc:hasMethod some loinc:*'Organism specific culture'*

C

$$\frac{\textit{Escherichia coli O157} \sqsubset \textit{Escherichia coli serotype}}{\textbf{\textit{hasComponent}} \textit{ 'Escherichia coli O157'} \sqsubset \textbf{\textit{hasComponent}} \textit{ 'Escherichia coli serotype'}}$$

Figure 5 Exemple de reclassification par surDéfinition et sousRestriction de tests formalisés avec le ontoLOINC. **A.** Définition du test subsumeur **B.** Définition du test subsumé **C.** Explication de l'inférence par sousRestriction

4 Résultats

4.1 Équivalence de tests

4.1.1 LOINC—SCT

La classification automatique de LOINC—SCT a mis en évidence l'équivalence de 45 tests deux à deux. Après une analyse de chacune des équivalences inférées nous observons que 70% s'expliquent par une différence de granularité entre les parties *Composant* et le concept SNOMED CT® utilisé pour les décrire. Ces résultats ont été présentés aux experts internationaux du domaine impliqués dans la création de cette ressource. Leur analyse met en évidence deux causes racines. Dans 44% des cas, l'équivalence entre tests met en avant un problème de la construction de la ressource LOINC—SCT résultant :

- de mauvais alignements entre partie LOINC® et concept SNOMED CT® (14 cas) ;
- de l'utilisation d'un concept défini pour représenter un test dont au moins une partie est alignée approximativement sur un concept SNOMED CT (6 cas).

Les autres équivalences mettent en exergue des problèmes terminologiques de LOINC. Les experts ont identifié 20 tests équivalents résultant d'une sémantique redondante et 6 tests équivalents causés par une description « non consistante ou floue » dans LOINC.

4.1.2 ontoLOINC

La classification automatique d'ontoLOINC a mis en évidence 3 702 relations d'équivalence (**owl:equivalentClass**) entre 4 516 tests distincts (6,3% des tests totaux). Nous pouvons observer deux causes d'équivalence. Tout d'abord 68% des relations d'équivalence résultent de l'absence de restriction portant sur la dimension *Composant* (2 518) pour les deux tests. Ces équivalences s'expliquent par l'absence de partie *Composant* dans le fichier TEST.TXT pour 2 626 tests. Dans 25% des cas (955), l'équivalence entre deux tests s'explique par la présence d'une information complémentaire à la dimension *Système* (appelée *Super Système*) qui n'est pas modélisée dans ontoLOINC.

Enfin 228 relations d'équivalence ne s'expliquent ni par l'absence de partie *Composant* ni par l'absence de modélisation du *Super Système*, dont 138 correspondent à des tests de la

classe MICRO. L'analyse de ces 138 relations a mis en évidence une cause d'équivalence. La distinction des tests équivalents est portée par la quatrième sous partie du champs *Composant* (McDonald *et al.*, 2017) ; celui-ci spécifie un numéro d'identification unique pour le test. Les tests *630-4 | Bacteria identified in Urine by Culture|* et *17972-1 | Bacteria **# 4** identified in Urine by Culture|* représentant un test de culture bactérienne sur un échantillon d'urine. Le test *17972-1* spécifie un numéro d'identification (information portée par #4). Ce sous composant est un artefact utilisé pour la structuration d'un compte rendu avec la syntaxe HL7 et n'apporte aucune information sémantique sur la nature du test.

Pour conclure, parmi les 3 702 relations d'équivalence, 93% montrent les limites de la construction de l'ontologie ontoLOINC que ce soit dans le processus d'extraction des données ou dans le modèle lui-même. Bien que nécessitant un certain nombre de corrections, la première version d'ontoLOINC propose cependant une formalisation correcte de plus de 90% des tests de la terminologie LOINC.

4.2 Réorganisation hiérarchique des tests

Dans cette section nous présentons les résultats de classification automatique des tests. Nous décrivons ces classification d'un point de vue structurel (table 2) et sémantique (figure 6). L'étude structurelle décrit la topologie des classifications. L'étude sémantique permet d'expliquer la classification en fonction des dimensions (éléments sémantiques, table 1) utilisées pour formaliser les tests. La figure 6 mesure la proportion d'inférences expliquées par chacune des 9 dimensions initiales de la terminologie.

4.2.1 LOINC—SCT

Table 2 Métriques des classifications pour ontoLOINC et LOINC—SCT

Information générales	ontoLOINC	LOINC—SCT
Nombre de tests décrits	70 795	13 756
Nombre de relations taxonomiques (non redondante)	25 859	7 961
Test subsumés (unique)	25 082 (35,4%)	6 789 (49,3%)
Topologie des arbres taxonomiques		
Profondeur médiane de reclassification	1 *(81,2%)*	1 *(61,8%)*
Profondeur maximale de reclassification	3 (0,7%)	4 (1,2%)
Nombre de tests poly hiérarchisés	777 (3,1%)	1 127 (16,7%)
Raison de classifications		
surDéfinition	25 838	2 580
sousRestriction	0	4 988
surDéfinition + sousRestriction	0	219
Non Expliqué	21	174

Le raisonneur ELK infère 7 961 relations hiérarchiques et permet de classer 49,3% des tests (6 789) représentés dans LOINC—SCT (Table 2). L'étude des raisons de classification montre d'une part que plus de 62% des classifications sont induites par sousRestriction. La figure 6 illustre l'impact de chaque dimension pour les classifications par sousRestriction de LOINC—SCT (violet pointillé). Nous observons que la classification des concepts utilisés pour décrire les parties *Composant* expliquent plus de 92% des classifications. De plus, la classification des concepts des parties *Systèmes*, *Méthode* et *Propriété* explique dans une moindre mesure (total <10%) les inférences dues à la sousRestriction. Nous observons également que 32% des classifications sont inférées par l'ajout d'une restriction dans la définition du test fils (surDéfinition). L'analyse des causes de surDéfinition (figure 6 tirets) montre que plus de 90% des surDéfinitions représentent l'ajout d'une restriction portant sur la

Méthode du test, les 10% restants sont dues à l'ajout d'une restriction représentant une dimension optionnelle (*Challenge*).

4.2.2 ontoLOINC

Le raisonneur ELK infère plus de 300 000 relations hiérarchiques dont 90% sont redondantes. Cette redondance s'explique par les 302 relations d'équivalence ; par la suite l'étude de la classification se fait sur les 25 859 relations non redondantes (Table 2). Tout d'abord, nous observons 39% des tests d'ontoLOINC ont été reclassés et uniquement par ajout d'une sous restriction (surDéfinition**)**. Ce résultat s'explique par l'absence de classification hiérarchique des parties dans ontoLOINC.

Parmi les classifications inférées, nous observons que 31,5% résultent de l'absence de parties *Composant* pour 2 626 tests. L'absence de classification par sousRestriction s'explique par l'absence d'organisation hiérarchique des parties (entre elles) dans ontoLOINC. La Figure 6 montre également que les classifications résultent principalement des dimensions *Méthode* (55,8%) et *Challenge* (32,5%). Enfin 1 959 reclassifications de tests sont dues à une surDéfinition de la dimension *System* (7,6%).

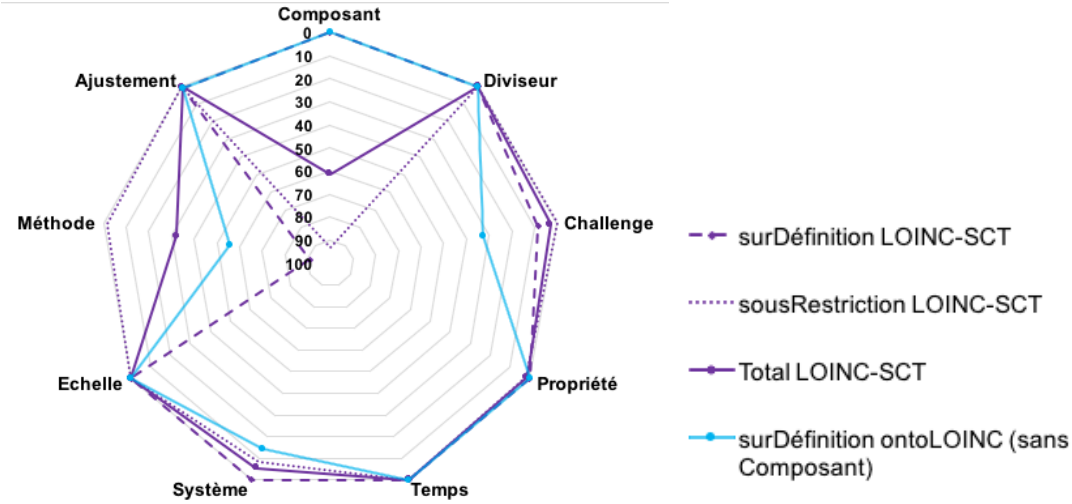


Figure 6 Impact des dimensions pour la reclassification par type de reclassification pour ontoLOINC (bleu) et LOINC—SCT (violet).

5 Discussion : des bénéfices d'une représentation ontologique

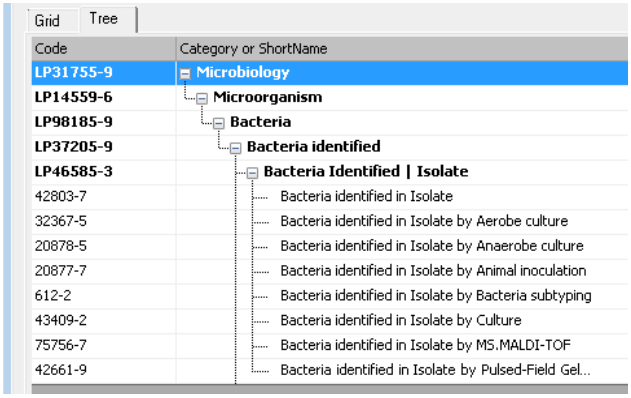
Dans cette section, nous présentons, dans un premier temps, l'impact d'une formalisation ontologique de LOINC® pour visualiser la terminologie. Dans une seconde section nous comparons les formalismes ontoLOINC et LOINC-SCT.

5.1 Classification des tests et visualisation

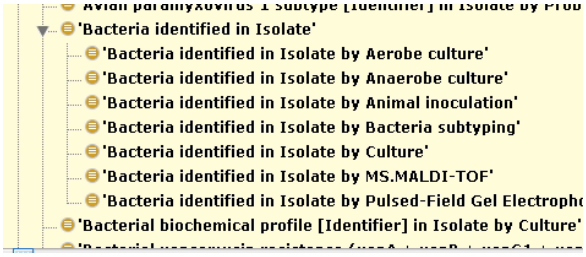
L'organisation des tests au sein d'une hiérarchie propose une visualisation contextuelle des tests les uns par rapport aux autres. La figure 7 montre les différences d'organisation hiérarchique autours du test *42803-7|Bacteria identified in Isolate|*. La terminologie LOINC® propose une visualisation des tests (hiérarchie Multi-Axiale - figure 7 A) sous forme de catégories construites en fonction des dimensions *Composant* et *Système ;* elle ne propose pas de classification des tests entre eux. ontoLOINC (figure 7 B) propose un premier niveau d'organisation des tests les uns par rapport aux autres. Dans l'exemple, on observe que la classification ontoLOINC affine la visualisation hiérarchique des tests en tenant compte des

parties *Méthode* (*Culture*, *Aerobic culture*, *etc.*). La classification obtenue avec LOINC-SCT (figure 7 C) illustre l'impact des sousRestrictions dans l'organisation hiérarchique des tests. On observe d'une part une organisation hiérarchique à trois niveaux, contre deux dans ontoLOINC, mais surtout l'apparition de nouveaux tests dans la hiérarchie. Ces nouveaux tests correspondent à des tests d'identification d'un sous ensemble de bactéries spécifiques, tel que les Streptocoques ou *Escherichia coli*. Cet exemple met en avant l'intérêt de la classification hiérarchique des parties pour enrichir la classification des tests.

A. Hiérarchie Multi-Axiale



B. ontoLOINC



C. LOINC-SCT

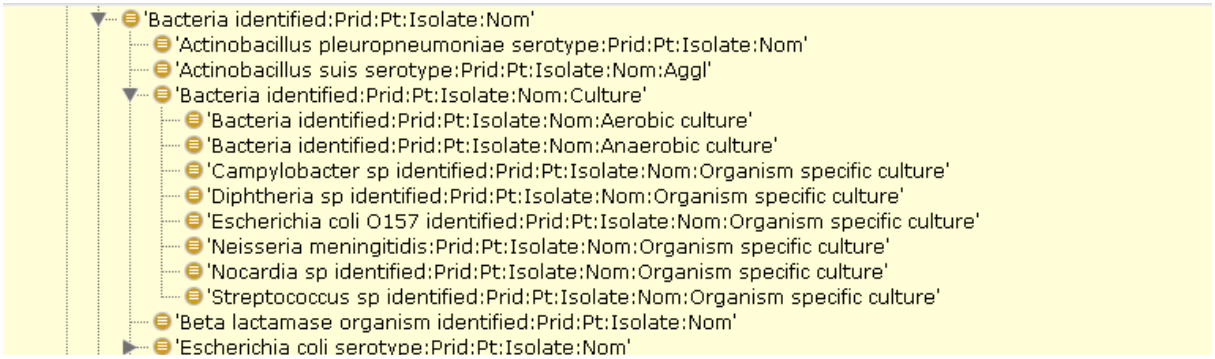


Figure 7 Organisation hiérarchiques autour du test *42803-7|Bacteria identified in Isolate|* proposé par la hiérarchie Multi-Axiale (A), et obtenus par classification automatique d'ontoLOINC (B) et LOINC-SCT (C).

5.2 Comparaison de l'organisation d'ontoLOINC avec LOINC—SCT

Si ontoLOINC et LOINC—SCT proposent une formalisation de la terminologie LOINC, leurs différences impacte leur utilisabilité.

LOINC—SCT est développée par une collaboration d'experts du *Regenstrief Institute* et de SNOMED International ce qui en fait une ressource de meilleure qualité qu'ontoLOINC. L'étude des équivalences a permis de mettre en évidence certains défaut de conceptions dans la terminologie. Contrairement à ontoLOINC elle propose un panel de classifications plus exhaustif grâce à la représentation hiérarchique des concepts utilisés pour formaliser les tests. Son utilisabilité est cependant limitée par trois inconvénients. D'une part son utilisation est restreinte par la licence d'exploitation de la SNOMED CT, ce qui complexifie le déploiement d'applications basées sur LOINC—SCT. D'autre part la version LOINC—SCT étudiée représente partiellement (20%) des tests de la terminologie LOINC. Si celle-ci est amenée à être incrémentée dans les prochaines version, seul les tests de laboratoires et les tests cliniques

Vital Sign ont vocation à être représenter avec SNOMED CT. Dans une étude précédente (Mary *et al.*, 2016), nous avons montré que l'utilisation de la ressource LOINC—SCT requière une compréhension fine de l'ontologie SNOMED CT pour exprimer des requêtes sur les tests.

Table 3 Comparaison des formalismes logiques de LOINC.

Caractéristiques Générales	**ontoLOINC (version 1)**	**LOINC—SCT (alpha release version 3)**
Libre de droit	Oui	Non
Version de la terminologie	Au choix (ici 2.52)	2.52
Couverture de la terminologie	Au choix (Test actif)	Partielle (20%)
Taille de l'ontologie	115 832 concepts	331 515 concepts
Créateur	Utilisatrice avertie de LOINC	Experts de LOINC® et SNOMED CT
Comparaison par rapport au modèle de la terminologie		
Isomorphisme	Oui	Non
Formalisation d'un test	1 patron explicitement décrit (modèle LOINC)	2 patrons décrits textuellement dans la documentation
Impact sur le développement et l'utilisation de la terminologie		
Identification de tests redondant	Non (trop de bruit)	Oui
Classification des parties	Pas dans la 1ère version	Oui

ontoLOINC, décrite dans ce papier, est une première version d'ontologie (mars 2017) qui nécessite encore quelques corrections. Nous sommes aujourd'hui en contact avec les membres du Regenstrief Institute pour identifier les meilleurs protocoles d'extraction des données de la dimension *Composant.* Le processus de construction de cette ontologie offre la possibilité de personnaliser de l'ontologie LOINC®. L'extraction des données permet de sélectionner les tests de n'importe quelle version de la terminologie, mais également sur d'autres critères comme la *classe* du test par exemple. Cependant l'utilisation de RELMA pour extraire les éléments nécessaires à la construction d'ontoLOINC nécessite une autorisation préalable du Regenstrief Institute (cf. Licence LOINC®).

Le patron de formalisation d'un test est strictement identique à la structure d'un test dans la terminologie LOINC® ; ce qui facilite l'assimilation de cette nouvelle ressource par des experts biologistes. Néanmoins l'absence d'organisation hiérarchique des concepts partie limite la reclassification des tests à des surDéfinitions.

6 Conclusion

Dans cet article nous présentons tout d'abord ontoLOINC, une ontologie permettant de représenter la terminologie LOINC® ainsi que son processus de génération. Nous avons ensuite étudié les bénéfices des formalismes logiques pour la classification des tests de laboratoire. L'étude de la classification de la ressource LOINC—SCT[4] (IHTSDO & Regenstrief Institute, 2013) a permis de démontrer les bénéfices d'une organisation hiérarchique des parties pour la classification automatique des tests et leur visualisation.

Nous travaillons actuellement à une seconde version d'ontoLOINC pour améliorer la formalisation de la terminologie LOINC® en ajoutant la dimension *Super Système* et corriger l'extraction des parties représentant la dimension *Composant*. Une fois corrigé et avec l'aval du *Regenstrief Intitute*, nous mettrons à disposition ontoLOINC sur les répertoires d'ontologies publiques tel que BioPortal[7].

Nous développons également un module d'intégration d'ontoLOINC avec d'autres ontologies ; celui-ci utilise des alignements simples pour affiner la classification des tests. Ce module d'intégration sera utilisé dans un premier temps pour compléter la comparaison

[7] https://bioportal.bioontology.org/

d'ontoLOINC et LOINC—SCT. Nous réutiliserons les alignements entre parties et concepts décrit dans la ressource LOINC—SCT pour générer une nouvelle ontologie LOINC® intégrée à SNOMED CT®. Celle-ci nous permettra de valider les performances du modèle ontoLOINC pour classer les tests par sousRestriction.

À moyen terme, nous souhaitons étudier les classifications de tests obtenus sur ontoLOINC intégrée avec d'autres ontologies reconnues comme standard en biologie :

— CheBI comme taxonomie des composés chimiques ;
— NCBI Taxonomie pour les tests d'identification ;

Pour conclure, cet article présente deux représentations formelles de la terminologie LOINC®. Ce mode de représentation offre de nouvelles opportunités pour manipuler les données de diagnostic *in vitro*. L'approche de représentation formelle de la terminologie LOINC® a déjà été exploitée pour construire de nouvelles applications, notamment pour sélectionner les données de diagnostic à fort intérêt épidémiologique (Eilbeck, *et al.*, 2013). La formalisation de LOINC® nous permet d'envisager des systèmes d'interprétation des résultats d'analyses de diagnostic *in vitro* comme par exemple l'interprétation des résultats de microbiologie (Bright, *et al.* , 2012).

Remerciement

Nous remercions l'ensemble du groupe de travail des *Observable Entity* de SNOMED International pour leur expertise concernant la ressource LOINC—SCT et plus particulièrement Suzanne Santamaria, Farzaneh Ashrafi, Swampna Abhyankar et Daniel Karlsson.

Référence

BHATTACHARYYA, S. B. (2016). Using SNOMED CT. Dans S. Bhattacharyya (dir.), *Introduction to SNOMED CT* (p. 157‑182). Singapore : Springer Singapore. doi:10.1007/978-981-287-895-3_9

BLUMENTHAL, D. (2010). Launching HITECH. *New England Journal of Medicine*, *362*(5), 382‑385. doi:10.1056/NEJMp0912825

BODENREIDER, O., SMITH, B., KUMAR, A., & BURGUN, A. (2007). Investigating subsumption in SNOMED CT: An exploration into large description logic-based biomedical terminologies. *Artificial intelligence in medicine*, *39*(3), 183–195.

BRIGHT, T. J., FURUYA, E. Y., KUPERMAN, G. J., CIMINO, J. J., & BAKKEN, S. (2012). Development and Evaluation of an Ontology for Guiding Appropriate Antibiotic Prescribing. *Journal of Biomedical Informatics*, *45*(1), 120‑128. doi:10.1016/j.jbi.2011.10.001

CORNET, R., & DE KEIZER, N. (2008). Forty years of SNOMED: a literature review. *BMC Medical Informatics and Decision Making*, *8*(Suppl 1), S2. doi:10.1186/1472-6947-8-S1-S2

EILBECK, K., JACOBS, J., MCGARVEY, S., VINION, C., & STAES, C. J. (2013). *Exploring the use of ontologies and automated reasoning to manage selection of reportable condition lab tests from LOINC*. Communication présentée au ICBO (p. 12–15).

IHTSDO, & REGENSTRIEF INSTITUTE. (2013). Regenstrief and the IHTSDO are working together to link LOINC and SNOMED CT. Repéré à https://loinc.org/collaboration/ihtsdo

KAZAKOV, Y., KRÖTZSCH, M., & SIMANCIK, F.. (2012). *ELK Reasoner: Architecture and Evaluation.*, Communication présentée au OWL Reasoner Evaluation Workshop (p. 12).

MACARY, F. (2007). IHDE, CDA et LOINC : des composants d'interopérabilité au service du partage des résultats de biologie médicale. *Spectra biologie*, *26*(158), 51‑57.

MARY, M., SOUALMIA, L. F., & GANSEL, X. (2016). *Projection des propriétés d'une ontologie pour la classification d'une ressource terminologique*. Communication présentée au Journée Francophones sur les Ontologies (p. 12), Bordeaux, France.

MARY, M., SOUALMIA, L. F., GANSEL, X., DARMONI, S. J., KARLSSON, D., & SCHULZ, S. (2017). *Ontological Representation of Laboratory Test Observables: Challenges and Perspectives in the SNOMED CT Observable Entity Model Adoption*. Communication présentée au AIME 2017 Proceedings (p. 10), Communication présentée au 16th Conference on Artificial Intellignece in Medicine, Vienna, Austria.

MCDONALD, C. J., HUFF, S., DECKARD, J., ARMSON, S., ABHYANKAR, S., & VREEMAN, D. J. (2017). *Loinc User Guide*. Indianapolis,USA : Regenstrief Institute.

MCDONALD, C. J., HUFF, S. M., SUICO, J. G., HILL, G., LEAVELLE, D., ALLER, R., FORREY, A., MERCER, KA., DEMOOR, G., HOOK, J., CASE, J. & MALONEY, P. (2003). LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clinical chemistry*, *49*(4), 624‑633.

SCHULZ, S., SUNTISRIVARAPORN, B., BAADER, F., & BOEKER, M. (2009). SNOMED reaching its adolescence: Ontologists' and logicians' health check. *International Journal of Medical Informatics*, *78*, *Supplement 1*, S86‑S94. doi:10.1016/j.ijmedinf.2008.06.004

SHEIDE, A., & WILSON, P. S. (2013). Reading up on LOINC. *Journal of AHIMA/American Health Information Management Association*, *84*(4), 58–60.

SNOMED INTERNATIONAL. (2017a). SNOMED CT Document Library. Repéré à https://confluence.ihtsdotools.org/display/DOC/SNOMED+CT+Document+Library

SNOMED INTERNATIONAL. (2017b). Observable and Investigation Model Project Home. Repéré à https://confluence.ihtsdotools.org/display/OBSERVABLE/Observable+and+Investigation+Model+Project+Home

STROETMAN, V., D. KALRA, P. LEWALLE, A. RECTOR, J. RODRIGUES, K. STROETMAN, G. SURJAN, B. USTUN, M. VIRTANEN, & P. ZANSTRA. (2009). *Semantic Interoperability for Better Health and Safer Healthcare* (Report Publication no KK-80-09-453-EN-C)(p1-34). European Commission.

VREEMAN, D. (2015). Guidelines for using LOINC and SNOMED CT Together. *Daniel Vreeman*. Repéré à https://danielvreeman.com/guidelines-for-using-loinc-and-snomed-ct-together-without-overlap/

Infarctus du myocarde : quelles sont les trajectoires de soins pronostiques du décès à l'hôpital?

Jessica Pinaire[1,2,3], Jerôme Azé[1], Sandra Bringay[1,4], Paul Landais[2,3]

[1] LIRMM, UMR 5506, Université Montpellier, France
prenom.nom@lirmm.fr

[2] BESPIM, CHU de Nîmes, France
Jessica.Pinaire@chu-nimes.fr

[3] ÉQUIPE D'ACCUEIL 2415, Institut Universitaire de Recherche Clinique, Université Montpellier, Montpellier, France
Paul.Landais@umontpellier.fr

[4] AMIS, Université Paul Valéry, Montpellier, France
Sandra.Bringay@univ-montp3.fr

Résumé : Les maladies cardiovasculaires représentent la première cause de mortalité dans le monde. En France, environ 120 000 personnes sont atteintes d'infarctus du myocarde par an ; 12 000 en décèdent lors de la crise, et 18 000 personnes en seront décédées un an après. Prévenir le risque de décès lié à l'infarctus du myocarde est un des objectifs que nous nous sommes fixés. Nous proposons une méthode pour identifier les parcours de soins les plus pronostiques du décès hospitalier. À partir des données médico-administratives issues du PMSI (Programme Médicalisé des Systèmes d'Information), nous extrayons des motifs séquentiels fréquents et nous les intégrons dans un processus de prédiction du décès par un score mesurant la similarité entre la trajectoire du patient et chacun des motifs extraits. Les résultats obtenus nous ont permis de mettre en évidence l'importance de la surveillance et du suivi de ces patients longtemps après leur infarctus.
Mots-clés : Infarctus du myocarde, PMSI, Trajectoires de patients, Fouille de données, Prédiction, Décès.

1 Introduction

Avec 17,5 millions de morts par an, les maladies cardiovasculaires représentent la première cause de mortalité dans le monde [1]. L'Organisation Mondiale de la Santé (OMS) estime que d'ici 2030 près de 24 millions de personnes mourront de maladies cardiovasculaires et ces affections demeureront la première cause de mortalité. Le risque majeur associé à l'infarctus du myocarde (IM) est le décès. En France, environ 120 000 personnes sont atteintes d'infarctus du myocarde par an ; 12 000 en décèdent lors de la crise, et 18 000 personnes en seront décédées un an après. Par ailleurs, les maladies cardiovasculaires constituent une part importante de la consommation des soins ; elles représentent le poste de dépenses le plus important de la consommation de soins et de biens médicaux. En France, les dépenses pour l'année 2002 concernant les maladies cardiovasculaires ont représenté 13,6% des dépenses publiques de santé soit 15,3 milliards d'euros (Heijink *et al.*, 2008). À mesure que la population vieillit, ces dépenses devraient augmenter considérablement (Heidenreich *et al.*, 2011).

Compte tenu de ces enjeux, de nombreux chercheurs universitaires s'intéressent à la consolidation et à l'enrichissement des connaissances médicales, mais aussi à la prévision des risques

1. http://www.who.int/cardiovascular_diseases/global-hearts/Global_hearts_initiative/en/

de mortalité associés aux maladies cardiovasculaires (Freemantle *et al.*, 2013; Fox *et al.*, 2006). Étant donné le nombre de patients impliqués et la quantité de données à exploiter, les chercheurs ont également utilisé des méthodes de fouille de données (Rajalakshmi & Dhenakaran, 2015; Austin *et al.*, 2012), parfois combinées à des méthodes statistiques plus classiques pour étudier des motifs ou évaluer le risque de mortalité. D'autres auteurs se sont intéressés aux données séquentielles pour construire des modèles prédictifs, notamment en se basant sur des règles d'association, pour prédire les divers cheminements du patient entre les différentes unités médicales (Dart *et al.*, 2003), ou la prochaine étape du traitement médicamenteux (Wright *et al.*, 2015). Enfin, des chercheurs ont développé des modèles pronostiques du décès à partir des données médico-administratives (Freemantle *et al.*, 2013; Aylin *et al.*, 2007) et ont obtenus des résultats similaires à ceux des données cliniques voire, dans certains cas, de meilleurs résultats.

Dans cet article, nous proposons d'identifier les parcours de soins les plus pronostiques du décès hospitalier. À partir des données médico-administratives issues du PMSI (Programme Médicalisé des Systèmes d'Information), nous avons extraits des motifs fréquents dans des sous-populations (ou contextes). Ces motifs sont particulièrement intéressants car ils sont facilement interprétables par les experts. Nous démontrons également leur puissance prédictive pour prédire le risque de décès.

Nous avons intégré ces motifs dans les modèles prédictifs à l'aide d'un score. Nous avons comparé entre-elles les méthodes prédictives les plus utilisées dans la littérature et notre approche donne, non seulement des résultats compétitifs avec ceux de la littérature, mais également un modèle interprétable par l'expert.

Dans la section 2, nous introduisons le vocabulaire spécifique au domaine de la recherche de motifs séquentiels et nous l'illustrons avec un exemple d'applications à partir des données du PMSI. Ensuite, nous présentons le processus d'extraction de motifs, puis le processus de prédiction dans la section 3. Dans la section 4, nous identifions les parcours les plus à risque pour deux contextes particuliers. Enfin, nous discutons des résultats obtenus et de notre méthode dans la section 5.

2 Définitions préliminaires

2.1 Motif séquentiel

Définition 1 (Itemset, séquence et sous-séquences d'évènements)
Soit $I = \{i_1, i_2..., i_k\}$ l'ensemble de tous les items. Un sous-ensemble de I est appelé un ***itemset****. Une* ***séquence*** *$s = \langle e_1 e_2 ... e_m \rangle$ est une liste ordonnée d'itemsets, où $e_i \subset I$ pour $1 \leq i \leq m$. Une séquence $s' = \langle r_1 r_2 ... r_p \rangle$ est une* ***sous-séquence*** *de s, s'il existe des entiers $1 \leq n_1 \leq n_2 \leq ... \leq n_p \leq m$, tels que $r_1 \subseteq e_{n_1}, r_2 \subseteq e_{n_2}, ..., r_m \subseteq e_{n_p}$.*

Définition 2 (Support)
Soit une base de séquences $B = \{s_1, ..., s_n\}$, le support de la séquence s, $Freq_B(s)$, est le nombre de séquences dans B ayant s comme sous-séquence.

Définition 3 (Motif séquentiel fréquent)
Une séquence s est fréquente et appelée motif séquentiel si son support est supérieur ou égal à un support minimum $k > 0$ fixé : $Freq_B(s) \geq k$.

2.2 Motif séquentiel contextuel

Définition 4 (Contexte)
*Un **contexte** c est une catégorie ou une modalité d'une variable (e.g. Homme ou Femme pour le sexe). L'ensemble de tous les contextes muni d'une relation d'ordre partiel, $\leq$, constitue la **hiérarchie des contextes** H. Les **contextes feuilles** de H sont appelés les **contextes minimaux**. A contrario, plus on remonte dans l'arborescence de H plus le contexte est dit **général**.*

Définition 5 (Motif séquentiel contextuel fréquent)
Soit c un contexte, H la hiérarchie des contextes et s une séquence. Une séquence s est fréquente dans un contexte c si son support dans c est supérieur ou égal à un support minimum $k > 0$ fixé : $Freq_c(s) \geq k$.

2.3 Exemple

Dans cet exemple, nous allons considérer les évènements (les séjours hospitaliers), représentés par les GHM (Groupes Homogènes de Malades), de 14 patients sur une période de 4 mois. Le temps est divisé en estampilles temporelles représentées par les mois. Supposons qu'à chaque mois, il ne peut se produire qu'un seul évènement (*i.e* un patient a un seul séjour par mois). Ces informations sont contenues dans la base de données présentée dans le tableau 1. Elle décrit différents GHM (05M13 : Douleurs thoraciques ; 05M06 : Angine de poitrine ; 05M16 : Athérosclérose coronarienne ; 05M04 : IM aigu ; 05M09 : Insuffisance cardiaque) associés au cours du temps par des professionnels de santé à des séjours de patients.

TABLE 1 – Mise en valeur du motif $\langle$(05M13)(05M06)$\rangle$ (en gras) soit le GHM 05M13 suivi du GHM 05M06. Ce motif est fréquent dans la base pour un support minimum de 50%.

Patients	Janvier	Février	Mars	Avril
P_1		**05M13**	05M04	**05M06**
P_2	**05M13**	05M09	**05M06**	
P_3	**05M13**	05M13		**05M06**
P_4	05M16	**05M13**	**05M06**	05M16
P_5	05M04	**05M13**	**05M06**	05M04
P_6		05M06		05M13
P_7		**05M13**	**05M06**	05M13
P_8	05M04	**05M13**	05M16	**05M06**
P_9		**05M13**	05M13	**05M06**
P_{10}		05M06	05M16	05M04
P_{11}		05M06	05M04	05M13
P_{12}	05M09	05M06	05M04	05M13
P_{13}	05M06	05M04	05M09	
P_{14}	05M06		05M13	05M09

Ces données sont **séquentielles** car elles présentent des évènements (les GHM) disposés suivant un ordre (le temps). Par exemple, pour le patient P_{12}, le GHM 05M09 associé en janvier, le GHM 05M06 a été associé au séjour de février, puis le GHM 05M04 associé au séjour de mars, enfin le GHM 05M13 associé au séjour d'avril. Une **sous-séquence** de la séquence du patient P_{12} est par exemple, la séquence $\langle$(05M09)(05M06)$\rangle$. Elle est également présente dans

la séquence du patient P_2 : son **support** est donc de 14% (2 sur 14) [2]. En examinant le tableau 1, nous constatons que le motif 05M13 suivi plus tard par 05M06 est vérifié par plus de 50% des patients (8 sur 14). En supposant que le professionnel de santé précise qu'il est intéressé par des GHM qui apparaissent dans au moins 50% des cas (support minimum) présents dans la base alors il s'avère que la sous-séquence $\langle$(05M13)(05M06)$\rangle$ est un **motif séquentiel fréquent**.

Jusqu'à présent, nous avons considéré la base comme un ensemble indivisible pour la recherche des motifs. Maintenant, nous allons prendre en compte les circonstances liées aux données : **les contextes**. Nous intégrons des informations supplémentaires, dans le tableau 2, qui associent à chaque patient son âge (*jeune* ou *âgé*) et son sexe (*homme* ou *femme*).

Ces informations contextuelles peuvent avoir une influence non négligeable sur la séquence d'évènements. Ainsi, l'extraction de motifs doit rendre cette influence perceptible pour l'utilisateur afin de lui offrir une vue contextualisée des données. Considérons maintenant la séquence $\langle$(05M13,05M06)$\rangle$ dans le tableau 2, nous constatons que :

- cette séquence de GHM est fréquente dans la population âgée (7 personnes âgées sur 8) mais pas dans la population jeune (seulement 1 personne sur 6) ;
- cette séquence de GHM demeure fréquente chez les âgés quel que soit leur sexe (5 hommes âgés sur 5 et 2 femmes âgées sur 3).

TABLE 2 – Mise en valeur du motif $\langle$(05M13)(05M06)$\rangle$ (en gras) avec les informations contextuelles sur l'âge et le sexe. Ce motif est spécifique aux personnes âgées. Une seule personne jeune est concernée.

Patients	**Age**	**Sexe**	**Janvier**	**Février**	**Mars**	**Avril**
P_1	âgé	homme		**05M13**	05M04	**05M06**
P_2	âgé	homme	**05M13**	05M09	**05M06**	
P_3	âgé	homme	**05M13**	05M13		**05M06**
P_4	âgé	homme	05M16	**05M13**	**05M06**	05M16
P_5	âgé	homme	05M04	**05M13**	**05M06**	05M04
P_6	âgé	femme		05M06		05M13
P_7	âgé	femme		**05M13**	**05M06**	05M13
P_8	âgé	femme	**05M13**	05M16	**05M06**	
P_9	jeune	homme		**05M13**	05M13	**05M06**
P_{10}	jeune	homme		05M06	05M16	05M04
P_{11}	jeune	homme		05M06	05M04	05M13
P_{12}	jeune	femme	05M09	05M06	05M04	05M13
P_{13}	jeune	femme	05M06	05M04	05M09	
P_{14}	jeune	femme	05M06		05M13	05M09

3 Protocole de prédiction

La première étape de notre protocole consiste à extraire puis à trier des motifs contextuels fréquents à partir des données issues du PMSI. L'étape suivante de notre protocole de prédiction a été construite à l'aide des recommandations établies pour élaborer des modèles prédictifs

2. Notons que dans notre cas les itemsets sont réduits à un item.

à des fins de pronostic ou de diagnostic : la méthode TRIPOD (Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) (Collins *et al.*, 2015).

Ces étapes sont plus précisément expliquées dans la suite de cette section et sont schématisées dans la figure 2. Les expérimentations ont été réalisées à l'aide du logiciel R version 3.3.1.

3.1 Étape 1. Extraction de motifs séquentiels contextuels

L'objectif de cette première étape est d'extraire des profils de parcours de soins fréquents pour l'IM qui prennent en compte des informations contextuelles fréquemment associées aux données séquentielles. Ainsi, nous identifions des profils de parcours de soins spécifiques d'une sous-population donnée.

Le processus de fouille, s'effectue de la façon suivante :

1. Nous sélectionnons les patients ayant eu un IM à l'aide d'une requête SQL sur la base PMSI. Pour chaque patient nous récupérons l'ensemble de ses séjours sur la période 2009 à 2014, excepté les séjours pour séances (*e.g.* radiothérapie, dialyses, chimiothérapie, *etc.*) et les prestations inter-établissements [3]. Selon les règles de codage du PMSI, ces séjours ont le même motif d'admission ;
2. Nous créons des sous-populations appelées contextes, à l'aide de covariables associées aux patients. Pour ce faire, nous avons pris en compte le genre (Homme/Femme), la classe d'âge du patient au moment de sa première apparition dans la période d'observation : -45 ans, 45-65 ans et +65 ans [4]. Enfin, nous avons retenu le nombre de séjours selon trois catégories : les 3-5 séjours, les 5-60 séjours et enfin les plus de 60 séjours [5]. Nous obtenons 18 contextes minimaux. La hiérarchie de nos contextes est représentée dans la figure 1 ;

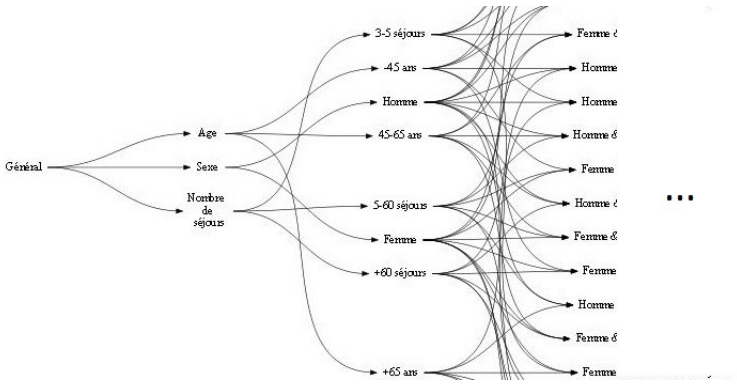


FIGURE 1 – Hiérarchie des contextes.

3. Nous construisons la base séquentielle en nous intéressant aux GHM, c'est-à-dire aux codes utilisés par les professionnels pour catégoriser un séjour et donnant lieu à tarifica-

3. Transfert d'un patient dans un autre établissement pour lui faire faire un acte (*e.g.* une coronarographie) car le premier n'a pas l'équipement pour le réaliser.

4. Ces classes d'âge ont été construites après concertation de l'expert médical.

5. Classes déterminées par l'expert médical : plus un patient est hospitalisé, plus il a des complications médicales associées à sa maladie.

tion. À chaque patient est associé une séquence de GHM, de longueur égale au nombre de séjours effectués pendant ces six années. Pour chaque patient, nous avons nettoyé automatiquement les données, en conservant la première hospitalisation du patient puis tous ses séjours d'hospitalisation liés à la cardiopathologie. Par exemple, pour un patient dont la séquence d'hospitalisation est : *IM, diabète, obésité, fracture du poignet et angine de poitrine*. Dans cette séquence d'évènements, nous identifions un séjour non lié à la cardiologie : *fracture du poignet*. Cet évènement est retiré de la séquence du patient. Les séquences de GHM ainsi triées constituent **la trajectoire du patient** ;

4. Nous effectuons la recherche de motifs séquentiels contextuels à l'aide de l'algorithme CFPM (Contextual Frequent Pattern Mining) (Rabatel, 2011) avec un support de 1%. Cela signifie que les motifs sont extraits s'ils concernent au moins 1% des patients d'un contexte ;
5. Nous procédons à la suppression de l'information que nous souhaitons prédire : le décès. Il s'agit des codes GHM contenant cette information de façon intrinsèque [6]. D'autre part, nous ne conservons que les motifs maximaux (Gouda & Zaki, 2005) non inclus dans un autre motif.

Nous passons ensuite à l'étape de prédiction.

3.2 Étape 2. Prédiction

Modéliser consiste à représenter un phénomène ou une situation observée afin de mieux l'étudier. Il s'agit généralement de trouver une représentation qui, à partir de paramètres, permet d'obtenir une retranscription qui soit la plus en adéquation possible avec les observations.

Ici, nous souhaitons prédire la mortalité hospitalière suivant le parcours du patient, ainsi la variable binaire à expliquer est l'état final du patient : présumé vivant ou décédé dans un établissement de soins.

Le processus de prédiction, s'effectue en 4 étapes :

1. Nous constituons des échantillons équilibrés [7] par contexte à partir des données issues du point 3 de l'étape 1, en ne conservant que les patients ayant eu au moins 4 évènements durant la période d'observation. Ceci nous permet de constituer une base de données avec des patients ayant un historique suffisant pour améliorer la capacité prédictive d'un modèle. De plus, nous écartons les patients originaires des régions Sud Méditerranée et Nord-Ouest [8] que nous conservons pour la validation externe [9] ;
2. Nous mesurons un score de similarité, entre les différents motifs et la trajectoire du patient. Nous intégrons la notion de distance dans le choix du modèle. Nous avons retenu les distances suivantes : la plus longue sous-chaîne commune (LCS), la distance de Levenshtein, la distance d'alignement optimal, la distance de Damerau-Levenshtein, les distances q-gramme, Jaccard, cosinus, Jaro et Jaro-Winckler ;

6. *e.g.* 05M21 signifiant Infarctus aigu du myocarde avec décès : séjours de moins de 2 jours.

7. L'équilibrage se fait sur la variable à prédire : autant de patients décédés que de patients vivants.

8. La région d'origine du patient est déterminée selon les territoires médicaux de la carte des inter-régions du CeNGEPS (`http://www.cengeps.fr/fr`).

9. Selon le principe d'une validation dite "géographique".

3. Nous prédisons avec plusieurs modèles : régression logistique (RL), Naïf Bayes (NB), k plus proches voisins (KNN), arbre de régression et Séparateur à Vastes Marges (SVM). Nous comparons les modèles et sélectionnons le meilleur selon les critères : accuracy, taux d'erreur, précision, F-mesure et aire sous la courbe ROC (Receiver operating characteristic (AUC)) ;
4. Nous validons le modèle sélectionné à partir de l'échantillon des patients ayant des trajectoires de longueur 4 et originaires des régions Sud Méditerranée et Nord-Ouest. Les critères d'évaluation sont : l'AUC et le score de Brier (Brier, 1950).

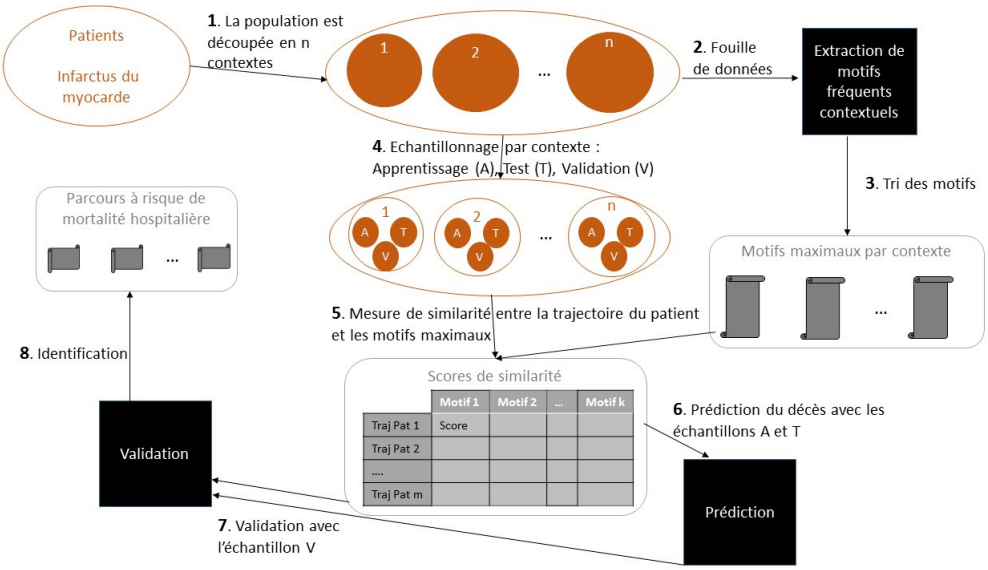


FIGURE 2 – Schéma du protocole de prédiction.

4 Expérimentations

Suite à l'application du protocole décrit dans la section 3, les modèles retenus sont ceux de la RL couplés à une distance d'édition. Dans cette section, nous identifions les parcours les plus à risque de la mortalité hospitalière pour les contextes suivants : les patients ayant 3-5 séjours et les patients ayant 5-60 séjours.

Pour identifier les parcours les plus à risque de la mortalité hospitalière, nous avons analysé l'influence des variables impliquées dans les modèles. Les résultats sont présentés dans le tableau 3.

Les résultats d'une RL sont interprétés en termes de facteurs de risque si l'OR est supérieur à 1, de facteurs protecteurs si l'OR est inférieur à 1 ou encore absence d'association entre l'évènement d'intérêt et la variable si l'OR est égal à 1.

Après examen de la première partie du tableau 3 nous identifions les motifs 05M04-05K06 (IM aigu suivi de pose de stent) et 05K10 (Actes diagnostiques par voie vasculaire) comme étant des parcours augmentant le risque du décès hospitalier pour le contexte 5-60 séjours. Tandis que les motifs 05K06-05K06-05K06 (3 séjours pour pose de stent) et 05K13 (Actes

TABLE 3 – Modèles logistiques pour les trajectoires de GHM.

		Modèle 5-60 séjours	
Variables	**Coefficient**	**OR**	**IC 95%**
Constante	-3,13***		
Scores			
05M04-05K06	9,63 **	1,52e+04	38,83 à 9,37e+06
05K10	43,08***	5,12e+18	5,13e+09 à 2,56e+28
05K06-05K06-05K06	-8,34 ***	2,39e-04	5,33e-06 à 0,009
05K13	-20,87*	8,62e-10	4,82e-17 à 0,007
		Modèle 3-5 séjours	
Variables	**Coefficient**	**OR**	**IC 95%**
Constante	-18,26***		
Classe d'âge			
45-65 ans	-2,26**	0,1	0,009 à 0,57
-45 ans	-2,96	0,05	3,89e-04 à 1,54
Scores			
05K10	137,38***	4,62e+59	7,37e+32 à 2,33e+101
05K13	-66,74**	1,03e-29	8,37e-56 à 6,41e-10
05M04	15,23	4,13e+06	4,24e-01 à 7,31e+14

Variables : variables retenues dans le modèle ;
Coefficient : valeur des β_i du modèle
et test de nullité des β_i avec *p<0,05 ; **p<0,01 ; ***p<0,001 ;
OR : Odds-ratio ; **IC 95%** : intervalle de confiance à 95% des OR.

thérapeutiques par voie vasculaire) sont des facteurs protecteurs. Pour le contexte 3-5 séjours, nous retrouvons des résultats assez similaires. Dans la deuxième partie du tableau 3, les motifs 05M04 et 05K10 sont des facteurs de risque du décès alors que le motif 05K13 est identifié comme un facteur protecteur. Par ailleurs, l'examen des résultats concernant les classes d'âge indique une augmentation du risque pour la classe des +65 ans.

5 Discussion

Dans cette section, nous discutons de la méthode et de ses performances, dans la partie 5.1, nous nous comparons à d'autres et nous envisageons d'autres outils. Ensuite, dans la partie 5.2, nous faisons une synthèse des résultats obtenus et nous argumentons sur les limites de l'interprétation de ces résultats.

5.1 Choix du modèle : performances et extensions possibles

La comparaison des différents modèles montre que la RL couplée à une distance d'édition obtient les meilleures performances dans la plupart des contextes. D'autres auteurs (Austin, 2007; Steyerberg *et al.*, 2001) obtiennent des résultats similaires concernant la compétitivité de la RL comparée à d'autres modèles dans le cas de la prédiction de la mortalité à 30 jours après un IM aigu. En outre, le seul cas où nous obtenons un modèle avec une distance q-gramme est le contexte des femmes avec 5-60 séjours. Or, de manière générale, les femmes ont des trajectoires plus courtes (dans notre sélection des données). Le choix de la distance est donc lié à la longueur des séquences.

Par ailleurs, une étude comparative (Siontis *et al.*, 2012) des travaux de modélisation du risque de la mortalité, dans le cas de maladies cardiovasculaires, réalisés à partir de données cliniques, montre que les performances selon l'AUC varient de 0,71 à 0,88. Or nous obtenons dans le cas de la sélection des meilleurs modèles des performances variant de 0,6 à 0,98. Nos modèles ont donc des performances comparables à ceux évoqués plus haut. En outre, nous constatons que les résultats sont meilleurs pour des contextes à faibles effectifs. En effet, dans ces contextes, l'échantillonnage arrive à recouvrir plus de situations diverses. Ainsi, les échantillons sont plus représentatifs de la population et de fait les modèles n'en sont que meilleurs.
Toutefois, nous pourrions affiner nos résultats en intervenant sur différentes étapes de notre protocole, notamment à l'aide de techniques pour la sélection de prédicteurs (Guyon & Elisseeff, 2003; Claeskens *et al.*, 2008). En outre, d'autres approches pourraient être envisagées pour étudier les trajectoires de patients, comme l'analyse formelle de concept utilisée dans (Jay *et al.*, 2013) pour obtenir une représentation hiérarchique de l'information. Cette dernière pourrait être combinée à une méthode de calcul d'évènements (Event Calculus) (Mueller, 2008) afin de déterminer les évènements liés au décès. Une autre approche possible est celle employée dans (Fabregue *et al.*, 2011). Les motifs sont classés selon l'état final du patient et utilisés comme descripteurs d'un classifieur classique pour déterminer de quel groupe (Vivant/Décédé) un patient est le plus proche en fonction de sa trajectoire.

5.2 Les parcours à risque : résultats et limites de l'interprétation

La modélisation du décès à l'aide des motifs fréquents permet de distinguer les évènements hospitaliers favorisant une augmentation du risque de décès de ceux qui, au contraire, ont un effet protecteur. D'après les résultats décrit dans la section 4, les motifs préservant du décès sont les actes thérapeutiques et le parcours IM aigu suivi d'une pose de stent. Ceci atteste de l'efficacité de la prise en charge (Falconnet *et al.*, 2009) avec un suivi des soins de la dilatation artérielle par divers moyens : endoprothèse, angioplastie... associés à un traitement médicamenteux. En revanche, suivant l'état de gravité de la pathologie, un acte exploratoire, comme une artériographie ou une coronarographie, représentant déjà un risque pour le patient (comme tout acte invasif), favorisera d'autant plus une augmentation du risque (Mottier & Baba-Ahmed, 2006). Ceci explique la présence de 05K10 dans les motifs à risque. Les motifs fréquents identifiés et intégrés dans un modèle prédictif du décès hospitalier viennent souligner l'importance du suivi des patients atteints de cette pathologie sur une période d'un an voire au-delà, comme en atteste d'ailleurs la littérature sur ce sujet (Neff, 2004).

D'autres auteurs ont également utilisé les bases médico-administratives pour de la prédiction. Par exemple, (Aylin *et al.*, 2007) ont comparé les modèles pronostiques du décès hospitalier à partir des bases administratives et des bases cliniques (registres nationaux vasculaires et cancers). Ils ont obtenu des résultats similaires avec les deux types de bases et ont ainsi démontré l'intérêt d'utiliser les bases administratives pour construire des modèles pronostiques. Un autre exemple d'étude est celui de (Jensen *et al.*, 2014) dans le cas des maladies chroniques afin de prévoir les flux de patients et d'envisager les infrastructures.

Néanmoins, nous pouvons formuler quelques critiques au regard de cette méthode, notamment dans le choix de la base de données. En effet, le PMSI est un outil d'allocation budgétaire, mais il a des limites dans le domaine épidémiologique car il expose à des imprécisions et à des erreurs tenant, entre autres raisons, à l'insuffisance de l'information, à des difficultés de

codage et à des nomenclatures inappropriées (adéquation de la codification de la maladie avec la réalité). Par voie de conséquence, la fiabilité du codage des séjours via les bases médico-administratives est controversée (Lombrail *et al.*, 1994). Pourtant, elles représentent indéniablement une source importante d'informations. Elles couvrent la majorité des établissements de santé privés et publics sur le plan national et recèlent de données médicales sur tous les séjours hospitaliers. Une alternative pour réduire ce bais intrinsèque au codage pourrait être l'appariement des informations avec les bases de données du Sniiram (Système nationale d'information inter-régimes de l'assurance maladie).

6 Conclusion

En utilisant les motifs séquentiels nous avons élaboré des modèles pour prédire le décès au sein d'un établissement de santé. Ces motifs ont été intégrés par des scores en mesurant la similarité entre la trajectoire du patient et les motifs. Le choix du score n'étant pas évident, nous avons choisi de mettre en concurrence différentes mesures entre chaînes de caractères de la littérature. Nous avons construit un protocole de prédiction qui s'articule en plusieurs étapes en nous appuyant sur la méthode TRIPOD. Nous avons introduit les modèles prédictifs les plus couramment employés afin de les comparer. *In fine*, notre objectif était double : 1) déterminer le couple *(modèle, score)* ayant les meilleures performances pour chacun des contextes ; 2) identifier les motifs favorisant une augmentation du risque de décès.

Notre avons atteint notre premier objectif. Il résulte de la comparaison entre les différents modèles que le modèle logistique couplé à une distance d'édition est le modèle offrant les meilleures performances avec la conservation des scores de similarités en variables continues. D'autres perspectives de comparaisons et de modélisation sont envisageables à l'aide des modèles de survie tels que Cox (Timsit *et al.*, 2005) ou encore des modèles prédictifs basés sur les séquences.

Pour affiner ces investigations, nous prévoyons d'employer d'autres types de motifs comme par exemple, les motifs obtenus avec la r-confiance (Mercadier *et al.*, 2016) qui permettent d'identifier les parcours les plus probables à partir d'un ou plusieurs premiers évènements. Nous souhaitons construire des modèles prédictifs en sélectionnant les parcours les plus représentatifs à la fois en fréquence, en taille et en confiance. Une autre approche pourrait également être envisagée, en tenant compte de l'état final du patient à la fin de la période d'observation dans la répartition des contextes. Cette approche consisterait alors à mettre en évidence des motifs qui soient spécifiques du décès. Ainsi, nous pourrions, soit intégrer ces motifs dans notre protocole de prédiction, soit mettre en œuvre une méthode de classification des patients à partir de leur trajectoire comme dans (Fabregue *et al.*, 2011) afin de prédire le décès.

Nous avons également atteint notre deuxième objectif en distinguant les motifs présentant un risque accru de décès. Ces motifs difficiles à interpréter par des experts médicaux, s'avèrent, en revanche utiles pour prédire le risque de mortalité hospitalière d'un patient. En effet, à l'issue de la fouille, les résultats obtenus mettent en évidence les motifs fréquents dans des sous-populations mais ne permettent pas à l'expert d'interpréter les résultats car ils n'ont pas de particularité populationnelle. L'implémentation de ces résultats dans un modèle favorise leur

interprétation dans la mesure où ils sont imputables à une cause : le décès. Nous retenons de nos expérimentations, présentées dans la partie 4, que le risque de décès est fortement influencé par le profil d'évolution de la maladie et le suivi du patient après IM. Ceci témoigne de l'importance des recommandations de la société française de cardiologie (Delahaye *et al.*, 2001) sur la surveillance régulière des patients après un IM au moins durant une année. En effet, le risque de rechute et de décès est encore très élevé durant cette période et même encore au-delà.

Références

AUSTIN P. C. (2007). A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Statistics in Medicine*, **26**(15), 2937–2957.

AUSTIN P. C., LEE D. S., STEYERBERG E. W. & TU J. V. (2012). Regression trees for predicting mortality in patients with cardiovascular disease : what improvement is achieved by using ensemble-based methods ? *Biometrical Journal*, **54**(5), 657–673.

AYLIN P., BOTTLE A. & MAJEED A. (2007). Use of administrative data or clinical databases as predictors of risk of death in hospital : comparison of models. *British Medical Journal*, **334**(7602), 1044–1052.

BRIER G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**(1), 1–3.

CLAESKENS G., CROUX C. & KERCKHOVEN J. V. (2008). An Information Criterion for Variable Selection in Support Vector Machines. *Journal of Machine Learning Research*, **9**(Mar), 541–558.

COLLINS G. S., REITSMA J. B., ALTMAN D. G. & MOONS K. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) : the TRIPOD statement. *BioMed Central Medicine*, **13**(1), g7594.

DART T., CUI Y., CHATELLIER G. & DEGOULET P. (2003). Analysis of hospitalised patient flows using data-mining. *Studies in Health Technology and Informatics*, **95**, 263–268.

DELAHAYE F., BORY M., COHEN A., DANCHIN N., DE GEVIGNEY G., DELLINGER A., FRABOULET J.-Y., GAYET J.-L., GUIZE L., IUNG P., MABO C., MONPÈRE P.-G., STEG D. & THOMAS (2001). Recommandations de la société française de cardiologie concernant la prise en charge de l'infarctus du myocarde après la phase aiguë. *Archives des maladies du cœur et des vaisseaux*, **94**(7), 697–738.

FABREGUE M., BRINGAY S., PONCELET P., TEISSEIRE M. & ORSETTI B. (2011). Mining microarray data to predict the histological grade of a breast cancer. *Journal of Biomedical Informatics*, **44**(Suppl. 1), S12–S16.

FALCONNET C., PERRENOUD J.-J., CARBALLO S., ROFFI M. & KELLER P.-F. (2009). Syndrome coronarien aigu : guidelines et spécificité gériatrique. *Revue médicale suisse*, **5**(204), 1137–1147.

FOX K. A. A., DABBOUS O. H., GOLDBERG R. J., PIEPER K. S., EAGLE K. A., WERF F. V. D., AVEZUM A., GOODMAN S. G., FLATHER M. D., ANDERSON F. A. & GRANGER C. B. (2006). Prediction of risk of death and myocardial infarction in the six months after presentation with acute coronary syndrome : prospective multinational observational study (GRACE). *British Medical Journal*, **333**(7578), 1091–1094.

FREEMANTLE N., RICHARDSON M., WOOD J., RAY D., KHOSLA S., SUN P. & PAGANO D. (2013). Can we update the Summary Hospital Mortality Index (SHMI) to make a useful measure of the quality of hospital care ? An observational study. *British Medical Journal Open*, **3**(1), e002018.

GOUDA K. & ZAKI M. J. (2005). GenMax : An Efficient Algorithm for Mining Maximal Frequent Itemsets. *Data Mining and Knowledge Discovery*, **11**(3), 223–242.

GUYON I. & ELISSEEFF A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, **3**(Mar), 1157–1182.

HEIDENREICH P. A., TROGDON J. G., KHAVJOU O. A., BUTLER J., DRACUP K., EZEKOWITZ M. D., FINKELSTEIN E. A., HONG Y., JOHNSTON S. C., KHERA A., LLOYD-JONES D. M., NELSON S. A., NICHOL G., ORENSTEIN D., WILSON P. W. F. & WOO Y. J. (2011). Forecasting the Future of Cardiovascular Disease in the United States. *Circulation*, **123**(8), 933–944.

HEIJINK R., NOETHEN M., RENAUD T., KOOPMANSCHAP M. & POLDER J. (2008). Cost of illness : An international comparison. *Health Policy*, **88**(1), 49–61.

JAY N., NUEMI G., GADREAU M. & QUANTIN C. (2013). A data mining approach for grouping and analyzing trajectories of care using claim data : the example of breast cancer. *BioMed Central Medical Informatics and Decision Making*, **13**(130).

JENSEN A. B., MOSELEY P. L., OPREA T. I., ELLESØE S. G., ERIKSSON R., SCHMOCK H., JENSEN P. B., JENSEN L. J. & BRUNAK S. (2014). Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Communications*, **5**, 4022.

LOMBRAIL P., MINVIELLE E., COMAR L. & GOTTOT S. (1994). Programme de médicalisation des systèmes d'information et épidémiologie : une liaison qui ne va pas de soi. *Revue d'épidémiologie et de santé publique*, **42**(4), 334–344.

MERCADIER Y., PINAIRE J., AZÉ J., BRINGAY S. & TEISSEIRE M. (2016). La r-confiance pour l'identification de trajectoires de patients. In *Proceedings des 16ème Journées Francophones Extraction et Gestion des Connaissances*, p. 535–536.

MOTTIER D. & BABA-AHMED M. (2006). Anticoagulants et gestes invasifs. *Médecine thérapeutique*, **12**(1), 48–52.

MUELLER E. T. (2008). Event calculus. *Foundations of Artificial Intelligence*, **3**, 671–708.

NEFF M. J. (2004). Practice Guidelines : ACC/AHA Release Guidelines on Management of Patients with STEMI : Hospital and Long-Term Management. *American Family Physician*, **70**(10), 2011–2021.

RABATEL J. (2011). *Extraction de motifs contextuels : Enjeux et applications dans les données séquentielles*. PhD thesis, Université Montpellier II.

RAJALAKSHMI K. & DHENAKARAN S. S. (2015). Analysis of Datamining Prediction Techniques in Healthcare Management System. *International Journal of Advanced Research in Computer Science and Software Engineering*, **5**(4), 1343–1347.

SIONTIS G. C. M., TZOULAKI I., SIONTIS K. C. & IOANNIDIS J. P. A. (2012). Comparisons of established risk prediction models for cardiovascular disease : systematic review. *British Medical Journal (Clinical research edition)*, **344**, e3318.

STEYERBERG E. W., EIJKEMANS M. J., HARRELL F. E. & HABBEMA J. D. (2001). Prognostic modeling with logistic regression analysis : in search of a sensible strategy in small data sets. *Medical Decision Making : An International Journal of the Society for Medical Decision Making*, **21**(1), 45–56.

TIMSIT J.-F., ALBERTI C. & CHEVRET S. (2005). Le modèle de Cox. *Revue des maladies respiratoires*, **22**(6), 1058–1064.

WRIGHT A. P., WRIGHT A. T., MCCOY A. B. & SITTIG D. F. (2015). The use of sequential pattern mining to predict next prescribed medications. *Journal of Biomedical Informatics*, **53**, 73–80.

Suivi et détection des idéations suicidaires dans les médias sociaux

Bilel Moulahi[1], Jérome Azé[1], Sandra Bringay[1,2]

[1] LIRMM, UNIVERSITÉ DE MONTPELLIER, CNRS
bilel.moulahi@lirmm.fr

[2] AMIS, Université Paul Valéry Montpellier

Résumé : L'utilisation croissante des médias sociaux permet un accès sans précédent aux comportements, aux pensées et aux sentiments des individus. Nous nous intéressons ici à l'évolution des états émotionnels des individus captés au travers des services de microblogging de type Twitter. Notre objectif est de prédire l'apparition d'idéations suicidaires. Dans ce travail, nous avons mis en place une chaîne de traitements permettant d'extraire des caractéristiques à partir des messages reflétant l'état émotionnel. Puis, nous appliquons un modèle basé sur les Conditionnal Random Fields pour prédire un nouvel état. L'originalité de l'approche est de prendre en compte l'historique des états émotionnels pour prédire le nouvel état. Une expérimentation préliminaire nous a permis d'évaluer notre approche sur des cas réels d'utilisateurs de Twitter. Ces type d'approche permet de mieux comprendre les liens entre expressions dans les médias sociaux et idéations suicidaires ainsi que les transitions entre états émotionnels.

Mots-clés : Média sociaux, Suicide, Conditionnal Random Fields.

1 Introduction

Les médias sociaux sont de plus en plus utilisés par les professionnels de santé pour détecter et diagnostiquer des troubles dépressifs majeurs (De Choudhury *et al.*, 2013; O'Dea *et al.*, 2015; Sueki, 2015; Adler *et al.*, 2016; Maigrot *et al.*, 2016). Les plateformes comme Twitter et Facebook facilitent l'auto-présentation sélective de comportements indésirables, tels que l'automutilation, l'anorexie, ainsi que l'expression d'émotions négatives liées aux idéations suicidaires, en particulier chez les jeunes.

Afin de mieux comprendre ces nouvelles pratiques, de nombreuses études comme celles de (Burnap *et al.*, 2015; Colombo *et al.*, 2016) ont porté sur la recherche dans le discours des individus des références à la dépression et aux idéations suicidaires. Des auteurs comme (Burnap *et al.*, 2015; Colombo *et al.*, 2016) soulignent une corrélation entre les taux de tentatives de suicide et le volume de messages liés aux idéations suicidaires publiés dans les médias sociaux. Si les médias sociaux peuvent affecter les individus en répandant des pensées suicidaires, ils peuvent aussi avoir un rôle positif en aidant ces individus à trouver du soutien moral. Par exemple, l'activité intense en ligne, notamment nocturne, est un signe précoce permettant d'anticiper une dégradation de l'état émotionnel d'un individu.

De méthodes efficaces sont désormais disponibles pour analyser les sentiments exprimés dans les réseaux sociaux (Barbosa & Feng, 2010; Kim *et al.*, 2013). Plusieurs études récentes tirent partie de ces travaux pour la détection et la surveillance des idéations suicidaires (Spasic *et al.*, 2012; Gunn & Lester, 2012; Poulin *et al.*, 2014) et d'états dépressifs (Moreno *et al.*, 2011; Karmen *et al.*, 2015). Toutefois, la plupart des méthodes de l'état de l'art prédisent les émotions véhiculées par les utilisateurs au niveau d'un message ou de l'individu. Elles ne prennent pas en compte l'évolution du comportement de l'individu. Or, les idées suicidaires sont incluses

dans un continuum de séquences d'événements influençant les états émotionnels et qui peuvent éventuellement conduire à une tentative de suicide (Adler *et al.*, 2016). Étant donné la nature séquentielle des contenus produits par les individus dans les médias sociaux, nous utilisons dans ces travaux un modèle basé sur les Conditional Random Fields (Lafferty *et al.*, 2001; Sutton & McCallum, 2012) qui permet de capturer l'évolution de l'état émotionnel des individus au fil du temps, en tenant compte du contexte passé et de l'activité du moment. Les état émotionnels sont au préalable inférés à partir des messages, via une analyse de textes permettant d'identifier les facteurs de risque (De Choudhury *et al.*, 2013).

Nous avons évalué notre approche sur un corpus de tweets annotés manuellement, publiés par des individus ayant exprimé des références au suicide. La collection d'individus a été validée par un psychiatre pour n'inclure que les utilisateurs ayant présenté de réels symptômes. Les résultats expérimentaux montrent que le système prédit correctement des séquences d'états mentaux.

Dans le reste de l'article, nous présentons un état de l'art succinct, puis décrivons notre approche pour la détection d'idéations suicidaires. Nous présentons et discutons les résultats, avant de conclure sur des perspectives.

2 Etat de l'art

En France, près de 10 000 personnes mettent fin à leurs jours chaque année, soit environ 25 par jour, selon le dernier rapport de l'OMS [1]. Deuxième cause de mortalité chez les 15-24 ans, après les accidents, le suicide est un fléau qui touche des adolescents souvent fragilisés par cette période charnière de la vie. Avec l'avènement des médias sociaux, les personnes à risque et notamment les jeunes, utilisent des outils comme Facebook, Twitter et Reddit pour exprimer des idéations suicidaires. Il est possible d'utiliser ces médias pour détecter de manière précoce les individus vulnérables et intervenir rapidement. En 2015, Facebook [2] a introduit un nouveau service permettant aux utilisateurs de rapporter un comportement suicidaire. Très récemment, ce service a évolué [3] pour permettre aux personnes qui visionnent un *live-stream* Facebook [4] d'interpeller son auteur ou de faire un signalement.

De nombreuses études ont porté sur les notes laissées par les individus avant un suicide. Ces notes ont été analysées en développant des classifieurs supervisés et non supervisés pour identifier les sujets discutés ainsi que les émotions exprimées par des personnes étant passées à l'acte (Spasic *et al.*, 2012; Pestian *et al.*, 2008).

Plus récemment, plusieurs études se sont intéressées à l'évaluation des facteurs de risque suicidaires dans les médias sociaux avec l'objectif de mieux comprendre ou de prévenir le suicide en détectant les idéations suicidaires de manière précoce. Par exemple, le projet Durkheim [5] étudie les activités des anciens combattants américains sur Twitter, Facebook et LinkedIn. L'objectif de ce projet est d'identifier les marqueurs de comportements à risque. Poulin *et al.* (2014) ont développé des modèles de prédiction en utilisant les textes des notes. Les résultats montrent

1. Observatoire national du suicide `http://www.who.int/topics/suicide/fr/`
2. `https://www.facebook.com/help/suicideprevention`
3. `https://newsroom.fb.com/news/2017/03/building-a-safer-community-with-new-suicide-prevention-tools/`
4. vidéo en direct
5. `http://www.durkheimproject.org/research/`

que les personnes qui se sont suicidées expriment souvent de la peur et une certaine agitation avant de passer à l'acte. Les modèles de prédiction proposés ont montré des taux d'exactitude proche de 65%. Gunn & Lester (2012) ont analysé les messages Twitter d'une jeune fille qui venait de se suicider, publiés les vingt-quatre heures précédant son décès. Ils ont trouvé une augmentation des émotions positives et un changement de la focalisation de soi à d'autres lorsque le moment du décès s'est approché. Les auteurs ont également étudié un éventail plus large de tweets. Pour cela, ils ont utilisé le logiciel Linguistic Inquiry and Word Count (LIWC) [6] pour identifier dans le discours, des mots porteurs d'émotions ainsi que des processus cognitifs. Sueki (2015) ont utilisé un panel en ligne de 250 jeunes d'une vingtaine d'années, utilisant régulièrement Twitter, pour examiner l'association entre les tweets liés au suicide et les passages à l'acte. Les auteurs ont étudié les caractéristiques linguistiques de l'idéation suicidaire et ont identifié les marqueurs les plus fréquents. Par exemple, des phrases comme "*I want to commit suicide*" sont fortement associées aux tentatives de suicide, alors que des phrases suggérant une intention suicidaire, comme "*I want to die*" y sont moins associés. Contrairement aux techniques populaires d'apprentissage, Karmen *et al.* (2015) ont combiné plusieurs méthodes de TAL pour filtrer les utilisateurs de forums et identifier les symptômes de dépression. Ces auteurs ont mis en correspondance les questionnaires traditionnels de dépistage de la dépression avec un ensemble de termes associés aux symptômes. Ensuite, ils détectent ces termes dans les textes et en déduisent un score au niveau du message.

La littérature actuelle manque de modèles efficaces pour prédire les tentatives de suicide. Actuellement, peu d'approches intègrent l'évolution du comportement de l'individu. L'analyse porte généralement sur un message ou sur l'ensemble des messages d'un individu. L'analyse ne permet pas de prédire à quel moment une personne présente un risque suicidaire. Maigrot *et al.* (2016) explorent une approche basée sur les concepts drift pour identifier un temps à risque. Une limite à leur approche est de ne pas expliciter les transitions entre les états émotionnels comme nous souhaitons le faire dans ce travail.

3 Un modèle basé sur le contexte pour le suivi et la détection les idéations suicidaires dans les médias sociaux

Dans cette section, nous reformulons le problème et décrivons le modèle utilisé pour monitorer les idéations suicidaires dans les médias sociaux. Étant donné une séquence de messages pouvant traiter de thèmes jugés à risque tels que la dépression, le suicide, l'automutilation ou l'anorexie, mais également contenir des thèmes sans rapport ou même des blagues dans un intervalle de temps très court, avec quelle précision pouvons-nous prédire qu'un individu présente un réel risque suicidaire ? Un modèle d'analyse de sentiments typique traite ce problème comme une tâche de classification multi-classes et prédit une étiquette pour chaque message indépendamment de la séquence entière. Dans ce travail, nous supposons que l'état émotionnel déduit d'un message au temps t, dépend des états émotionnels précédents. Notre hypothèse principale est que les états émotionnels peuvent être modélisés comme des observations dépendantes et continues, qui peuvent être capturées via des méthodes de traitement automatique de la langue puis prédites. L'état émotionnel peut être représenté soit par un état positif, neutre ou négatif (Barbosa & Feng, 2010), soit par des modèles plus complexes incluant des émotions comme la

6. https://liwc.wpengine.com

tristesse, l'espoir, l'excitation, etc. (Larsen & Diener, 1992; Yik *et al.*, 2011; Kim *et al.*, 2013). Dans la suite, nous considérons trois niveaux d'états émotionnels mais notre approche peut être facilement généralisée indépendamment du nombre d'états émotionnels.

3.1 Description du problème

Soit P = <p_1, p_2, ..., p_n> un flux continu de messages (tweets, messages facebook, etc.) ordonnés dans le temps dans une fenêtre temporelle W. Le problème consiste à prédire un vecteur Y = <y_1, y_1, ..., y_n> d'états émotionnels associés à la séquence de messages observée P. Les observations P en entrée sont représentées par des vecteurs d'attributs. Chaque observation p_j contient différentes informations à propos du message au temps t_j. Chaque variable y_j est un état émotionnel inféré à partir de l'observation p_j. La Figure 1 décrit une séries de messages impliquant un changement de l'état émotionnel.

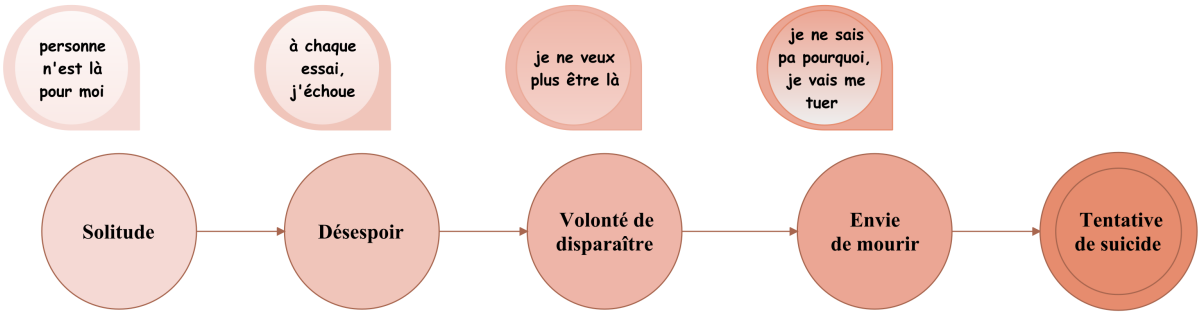


FIGURE 1 – Exemple d'évolution de l'état émotionnel d'un individu.

Ces changements d'états sont inspirés par un travail récent en psychologie cognitive (Adler *et al.*, 2016). Les auteurs ont étudié différents types de comportements suicidaires en explorant des signes cognitifs et les changements de comportements pouvant indiquer une tentative suicide. Par exemple, comme le montre la figure 1, un individu évoque sa *volonté de disparaître* avec le message "*Je ne veux plus être là*". Il est susceptible de voir son état émotionnel évoluer vers un niveau plus risqué comme *Envie de mourir* qu'il exprime avec le message "*Je vais me tuer*", avant de réaliser une *tentative de suicide*. Chaque nœud de la figure représente un état émotionnel qui est calculé en fonction d'une observation, le message de l'utilisateur. Les transitions (arêtes) entre les états encodent la séquence des changements d'état. Les modèles de réseaux de Markov permettent de représenter ces changements émotionnels séquentiels. Une généralisation intéressante est donnée par des modèles graphiques tels que les Conditional Random Fields (CRF) (Lafferty *et al.*, 2001).

Dans la suite, nous présentons les deux étapes de notre méthode. La première étape permet d'associer un ou plusieurs états émotionnels à un message à partir de l'analyse du texte de ce message. La deuxième étape permet de modéliser les interactions entre états émotionnels et peut être utilisée pour prédire le prochain état émotionnel d'un individu.

3.2 Extraction des caractéristiques

Avant d'extraire des caractéristiques des messages, ces derniers sont prétraités (mise en minuscules, suppression des ponctuations multiples, des caractères spéciaux, des mentions à d'autres utilisateurs et des URL). Nous avons extrait des caractéristiques largement utilisées dans la littérature (Spasic *et al.*, 2012).

– Le premier ensemble de caractéristiques inclut les caractéristiques lexicales du texte. Nous utilisons les étiquettes grammaticales (POS) pour capturer l'auto-référence (pronoms personnels à la première personne "*I*", "*My*", "*Je*", "*Mon*", etc), les noms, les verbes et les adverbes. De Choudhury *et al.* (2013) ont montré que l'utilisation de la première personne au singulier ou au pluriel peut révéler le bien-être ou le mal-être mental. Nous prenons également en compte l'intensité des émotions, en considérant la présence d'*intensifieurs* comme *très, complètement, intensément* surtout lorsqu'ils sont utilisés avec des pronoms personnels (eg., "je suis très triste"). Quand une phrase contient une négation suivie d'un symptôme, nous inversons cette caractéristique (eg., "je ne vais pas bien !").
– Le second ensemble de caractéristiques est lié aux lexiques. Nous cherchons dans les messages des termes couramment utilisés par les personnes à risque dans les médias sociaux. Nous considérons la fréquence des termes faisant référence aux émotions négatives, à la dépression, à l'automutilation, à la tristesse, à la santé mentale et au suicide. Pour ce faire, nous nous sommes inspirés des travaux de De Choudhury *et al.* (2013), qui ont exploité le lexique ANEW Bradley & Lang (1999) contenant un classement d'émotions pour un large nombre de mots [7]. Ensuite, nous enrichissons ces caractéristiques en incluant un autre lexique qui se réfère aux mots d'injure. En effet, De Choudhury *et al.* (2013) ont montré que ces caractéristiques véhiculent des informations importantes dans le contexte de l'analyse des états émotionnels.

3.3 Modèle basé sur les Conditional Random Fields

CRF est un type de modèle graphique probabiliste non dirigé qui a été appliqué avec succès dans de nombreux problèmes de traitement de textes et de visualisation (Sutton & McCallum, 2012). Un avantage de ce modèle réside dans sa capacité à capturer les dépendances complexes entre les observations, en plus des interprétations complètes de la relation entre les caractéristiques qu'il fournit. Dans notre contexte, cette propriété est très importante étant donné que la transition d'un état émotionnel à un autre dépend fortement des états observés précédemment. Par exemple, comme représenté dans la figure 1, il est très improbable que l'état émotionnel d'un individu saute soudainement de *solitude* (non risqué) à un état *tentative de suicide* (très risqué). La modélisation CRF est un modèle puissant qui aide à apprendre les comportements des utilisateurs et à prédire la séquence des états mentaux des utilisateurs.

Soit une séquence de messages observée $P = <p_1, p_2, ..., p_n>$ et une séquence d'états émotionnels cachés $Y =< y_1, y_1, ..., y_n>$, CRF modélise la probabilité conditionnelle comme suit :

$$p(Y|P) = \frac{1}{Z(P)} exp(\sum_{i=1}^{n} \sum_{k=1}^{F} w_k f_k(y_{i-1}, y_i, P, i)) \quad (1)$$

7. Ce lien contient le code et les scripts utilisés pour générer ce lexique : `https://github.com/sbma44/begin_anew`.

ou Z est un facteur de normalisation (aussi appelé la fonction de partition) pour que $p(Y|P)$ soit une probabilité valide pour toutes les séquences étiquetées. Z est défini comme la somme de l'exponentielle du nombre de séquences :

$$Z(P) = \sum_{P} exp(\sum_{i=1}^{n} \sum_{k=1}^{F} w_k f_k(y_{i-1}, y_i, P, i)) \quad (2)$$

Les paramètres w_k sont les poids des caractéristiques f_i. Ils sont appris par des techniques d'optimisation comme les approches par gradient. Les fonctions caractéristiques $f_k(y_{i-1}, y_i, P, n)$ prennent en compte une paire d'états émotionnels adjacents y_{i-1}, y_i, la séquence entière de messages P et la position courante dans la séquence i.

Notons que l'utilisation de CRF nous permet de définir un grand nombre de fonctions dépendantes ou indépendantes sans nous soucier de la relation statistique complexe entre ces fonctions. L'utilisation de chaque fonction dépend du poids w_k qui agit comme facteur d'activation de la fonction.

4 Expérimentations

Dans ce qui suit, nous détaillons la préparation du jeu de données utilisé, puis nous analysons les performances de l'approche en utilisant les caractéristiques détaillées dans 3.2. Nous explorons aussi l'importance des caractéristiques extraites à partir des messages et nous montrons les thèmes importants pour chaque état émotionnel en utilisant le modèle Latent Dirichlet Allocation Hoffman *et al.* (2010).

4.1 Préparation des données

En raison de l'absence de base de données librement accessible pour l'évaluation des méthodes de détection des risques de suicide dans les médias sociaux, nous avons utilisé l'API en temps réel Twitter [8] pour collecter des tweets contenant des références à des thèmes tels que la dépression, l'automutilation, l'anorexie et le suicide. La liste de mots clés utilisée pour récupérer les tweets a été définie manuellement à partir de la liste des facteurs de risque définie par l'APA (American Psychological Association [9]) et la liste des signes avant-coureurs définie par l'AAS (American Association of Suicidology [10]).

Parmi les tweets recueillis, nous n'avons conservé que ceux pour lesquels un symptôme lié au suicide a été validé par un psychiatre. 60 individus ont ainsi été choisis. Pour éviter un ajustement excessif du modèle, nous avons également inclus 60 comptes Twitter d'individus non à risque en utilisant les mêmes mots clés. Nous avons sélectionné au hasard les 50 derniers tweets de chacun de ces groupes. La collection totale de données contient 5976 tweets. Huit chercheurs et un psychiatre ont manuellement annoté 507 tweets de la collection pour déterminer les états émotionnels associés. Pour cette étude préliminaire, nous avons considéré trois états émotionnels. Le choix de ces classes est motivé par les travaux de Lehrman *et al.* (2012). Ces classes sont définies comme suit :

8. `https://dev.twitter.com/streaming/overview`
9. `http://www.apa.org/topics/suicide/`
10. `http://www.suicidology.org`

- *Aucune détresse* : le message traite d'événements quotidiens tels que le travail, les sorties, les activités du week-end, etc.
- *Détresse minimale/modéré* : le message exprime un niveau de détresse qui pourrait être considéré comme commun pour la plupart des individus tels que un examen, une présentation pour le travail, une dispute avec un ami, etc.
- *Détresse importante* : le message mentionne des références à l'auto-mutilation, aux idéations suicidaires, des excuses, des sentiments négatifs comme l'inutilité, la haine de soi, la culpabilité, etc.

Chaque tweet a été examiné par au moins deux annotateurs, avec un sous-ensemble de 55 tweets validés par le psychiatre. Nous avons calculé un kappa de Cohen de 69,1%, qui souligne un accord substantiel entre les annotateurs. Nous avons également calculé un kappa pondérée, qui tient compte des différents niveaux de désaccord, de l'ordre de 71,5% qui est un taux largement satisfaisant pour juger l'accord entre les annotateurs. Le processus d'annotation a donné 141 instances de la classe *aucune détresse*, 110 instances de la classe *détresse minimale* et 256 instances de la classe *détresse sévère*.

4.2 Protocole d'évaluation

Afin d'évaluer notre approche, nous avons adopté une méthodologie entièrement automatisée basée sur une validation croisée (k=5) sur l'ensemble des données annotés afin d'apprendre et tester le modèle proposé. Pour ce faire, à chaque itération, nous avons partitionné l'ensemble des 507 tweets annotés en échantillons d'apprentissage (70%) et de test (30%). Chaque instance est constituée par le tweet d'un utilisateur avec comme contexte l'ensemble des tweets formant la séquence des publications de l'utilisateur. L'objectif principal de la phase d'apprentissage est d'apprendre les paramètres de notre modèle ainsi que ceux des méthodes de référence (*baselines*). Nous comparons notre approche avec les deux modèles SVM et Random Forest en utilisant les mêmes ensembles d'apprentissage et de test. Nous avons utilisé les mesures d'évaluation : Rappel, Précision et F-mesure.

4.3 Résultat et discussion

4.3.1 Analyse des caractéristiques

Nous avons exploré l'importance des caractéristiques extraites à partir des messages en fonction des trois états émotionnels considérés dans la figure 2. Cette analyse a été effectuée sur l'ensemble des données d'apprentissage. Il est à noter que cette analyse est exploratoire, et nous ne l'avons pas utilisée pour la sélection des attributs les plus pertinents.

Chaque poids donne l'importance de la caractéristique pour l'état émotionnel. Alors que les poids élevés (positifs ou négatifs) indiquent une association forte, des pondérations nulles indiquent que la caractéristique a peu ou pas d'impact sur l'état émotionnel.

Dans ce travail, nous avons considéré 39 caractéristiques pour chaque message, mais pour des raisons de présentation, nous ne rapportons ici que les caractéristiques les plus importantes. Les *pronoms à la première personne* (lexical), les *mots d'injures* (lexique), les *Symptômes* (lexique) et *Symptômes antérieurs* (lexique) ont un impact négatif sur l'état émotionnel "Aucune détresse". En effet, les individus les moins à risque n'utilisent généralement pas de

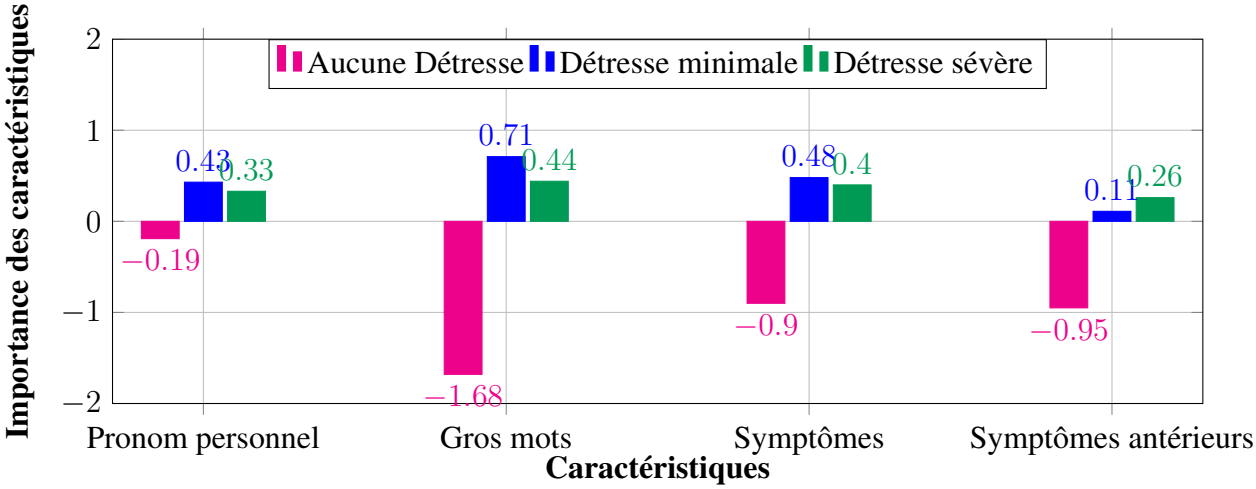


FIGURE 2 – Importance des caractéristiques dans les trois classes.

références aux symptômes ni de mots d'injures dans leurs messages. Au contraire, les injures et les symptômes sont des marqueurs importants pour la classe *détresse minimale* avec une légère différence pour la classe *détresse sévère*. D'autre part, la caractéristique *Symptômes antérieurs* qui reflète les symptômes antérieurs observés, est importante pour la classe *détresse sévère*. En effet, si à l'instant *t-1*, si un individu publie un message ayant une valeur importante pour la caractéristique *Symptôme*, il est probable que l'utilisateur présente une détresse émotionnelle sévère au moment *t*. La possibilité de prendre en compte ce dernier point est un avantage clé des modèles CRF.

Les tableaux 1 et 2 montrent les thèmes extraits des tweets appartenant aux états émotionnels *Aucune détresse* et *Détresse sévère*, en utilisant le modèle Latent Dirichlet Allocation (Hoffman *et al.*, 2010). Par souci de simplicité, nous avons fixé le nombre de thèmes à 2. Dans le tableau 1, les messages de l'état émotionnel *Aucune détresse* portent clairement sur la *famille*, les *voyages*, les *relations*, etc. alors que dans la Table 2, les thèmes abordés sont liés aux pensées suicidaires (e.g. suicide, meurtre, etc.). On remarque l'importance des intensificateurs appliqués aux termes reflétant des idées liées à la fin de vie (e.g. *assez, sans valeur, plus, fin*). Les figures 3 et 4 présentent les termes les plus utilisés pour les deux états émotionnels. Ces dernières corroborent les conclusions du modèle LDA.

4.3.2 Analyse des changements d'états émotionnels

Pour mieux comprendre les changements de comportement des utilisateurs, nous exploitons la puissance des CRF pour analyser les changements entre les états émotionnels. La figure 5 montre les transitions entre les 3 états émotionnels que nous avons considérés, en se basant sur l'ensemble des données d'apprentissage. Cette figure 5 permet d'identifier que les transitions les plus probables entre deux états différents vont de la classe *Aucune détresse* à la classe *Détresse minimale*, avec une probabilité inférieure pour la transition opposée. Les individus passant à un état émotionnel plus risqué sont peu susceptibles de revenir à un état normal. Les utilisateurs dans l'état *Aucune détresse* et *Détresse sévère* tendent à rester dans le même état avec des valeurs

Thème	Description
1	crazy tonight today morning hot **game** road **excitingtimes** haftflmdic **family-travel** arkansas cards poolside be life want up if day good everyone my yourself blessed hidline **girlideas** really **lol** know finally
2	**love** my me **happy** im bed **relationships** can always new better we few little snapchat going wait **excited** school suck drift tired makes exactly pat older attractive **dearly travel** catch

TABLE 1 – Thèmes extraits des tweets de la classe *Aucune détresse*.

FIGURE 3 – Aucune Détresse.

Thème	Description
1	me if would up one really **fucking** will **no** everyone am give **enough** days **everything** good **suicidal** still my seems **suicide** like **never** be **fuck** alive **die** say **help sleep**
2	**want myself tired** can my don **anymore** be wish not **kill worthless** never **feel die** will am **hate** end time **sorry** much me like **enough** live everything many **fight depressionproblems**

TABLE 2 – Thèmes extraits des tweets de la classe *Détresse sévère*.

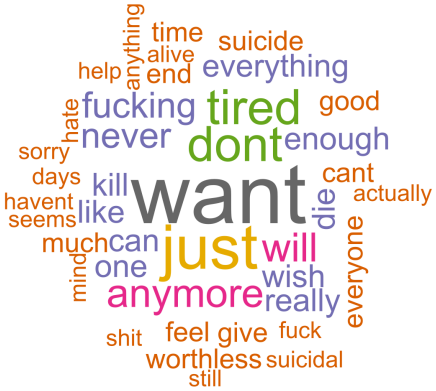


FIGURE 4 – Détresse sévère.

de 0,54 et 1,41, respectivement. La valeur de probabilité très faible pour la transition de *Détresse sévère* à *Détresse minimale* s'explique car l'état émotionnel *Détresse sévère* est généralement atteint lorsque l'individu se focalise sur une possible tentative de suicide (Adler *et al.*, 2016).

La Table 3 fournit les résultats des mesures de précision, de rappel et le F1-score pour toutes les classes, ainsi qu'une comparaison avec les deux modèles d'apprentissage automatique SVM et Random Forest. Dans la phase d'apprentissage, ces modèles sont construits en utilisant les paramètres par défaut de WEKA [11]. Pour apprendre notre modèle CRF, nous exploitons une descente de gradient en utilisant la méthode L-BFGS. Les coefficient de régularisation L1 et L2 du modèle sont fixés par défaut à 0.1 et 0.1, respectivement.
Les meilleurs résultats sont ceux des classes *Aucune détresse* et *Détresse sévère* selon le score F1, en dépit de la valeur importante de précision pour l'état émotionnel *Détresse minimale*. En

11. www.cs.waikato.ac.nz/ml/weka/

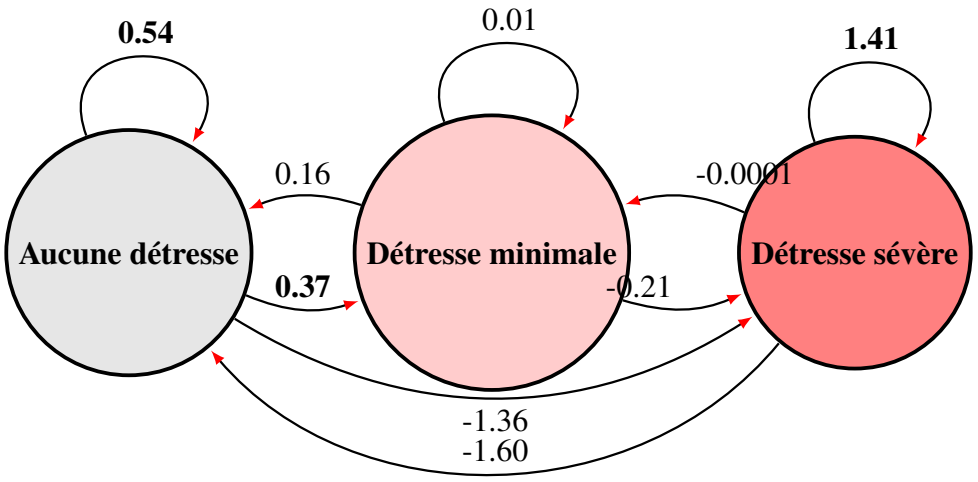


FIGURE 5 – Changement des émotions des utilisateurs selon les trois classes. Les valeurs en gras sont les plus probables. La couleur symbolise la sévérité de l'état émotionnel.

effet, le taux élevé de faux négatifs pour l'état de détresse minimale peut s'expliquer par le fait que : (i) cet état se trouve entre deux états ; (ii) l'auto-transition pour cet état est très faible (0.01) par rapport aux transitions entrantes et sortantes, en particulier depuis l'état *Aucune détresse*. D'autre part, le modèle permet d'identifier tous les messages appartenant à l'état *Aucune détresse*, ce qui n'est pas surprenant compte tenu de l'analyse de l'importance des caractéristique présentée dans la figure 2 (cf. section 4.3.1). Les résultats présentés dans le tableau 3 sont prometteurs car la plupart des méthodes de classification des textes pour des applications liées au suicide ou à la dépression atteignent à peine 0,7 (Burnap *et al.*, 2015; O'Dea *et al.*, 2015). Les résultats obtenus par les méthodes d'apprentissage automatique sont moins importants avec une grande différence en comparaison avec Random Forest. Les valeurs sont plus faibles en termes de Rappel pour les deux méthodes. Cette différence de performance peut être expliquée par l'absence de sélection d'attributs qui pourrait être considéré comme un désavantage dans les tâches de classification de texte.

	Précision	**Rappel**	**F1-score**
Aucune détresse	0.706	**1.000**	0.828
Détresse minimale	**1.000**	0.176	0.300
Détresse sévère	**0.941**	0.571	0.711
Notre approche	**0.816**	**0.752**	**0.711**
SVM	0.446	0.227	0.301
Random Forest	0.500	0.127	0.202

TABLE 3 – Évaluation des résultats du système de monitoring.

5 Conclusion et Perspectives

Dans cet article, nous avons proposé une approche pour le dépistage des idéations suicidaires basée sur un modèle probabiliste appelé Conditional Random Fields qui permet de modéliser et prédire le comportement en ligne de l'individu comme une séquence d'états émotionnels évoluant au fil du temps. Cette représentation permet d'incorporer un ensemble riche de caractéristiques complexes intégrant le contexte des messages précédents. L'efficacité de l'approche a été évaluée sur des données réelles, sur une collection de tweets publiés par des individus ayant montré des symptômes graves liés au suicide. Ces évaluations préliminaires ont montré que notre modèle est capable de fournir des interprétations complètes de la relation entre les états émotionnels et les résultats en termes de prédictions sont encourageants par rapport à la littérature.

Un avantage de l'approche est que nous pouvons facilement incorporer de nouvelles caractéristiques liées au texte en incluant notamment de nouveaux lexiques mais également non liées aux textes comme des informations contextuelles : l'heure de la rédaction du message, sa longueur, etc mais encore des caractéristiques liées à d'autres médias (images, vidéos...) associés aux messages. En effet, le modèle CRF permet d'incorporer un ensemble riche de caractéristiques représentant le contexte sans se soucier de leurs relations a priori (corrélations positives ou négatives). Cette flexibilité nous permet d'intégrer dans le modèle un ensemble de caractéristiques dont les dépendances peuvent être assez complexes et mal connues.

Un point important consiste à filtrer les références réelles liées au suicide par rapport aux messages de support et de condoléances, ou encore les campagnes de prévention du suicide (Burnap *et al.*, 2015). Par ailleurs, nous pouvons intégrer une représentation plus complexe de l'état émotionnel que seulement trois états. Par exemple, dans une étude clinique qui a été menée sur des jeunes étudiants, Moreno *et al.* (2011) ont établi un ensemble de critères cliniques de suicide qui peuvent être présents dans les publications Facebook. Nous citons à titre d'exemple la dépression, perte d'intérêt/plaisir dans les activités, changements d'appétit, problèmes de sommeil, agitation psychomotrice ou retard, perte d'énergie, sentiment d'inutilité ou de culpabilité, diminution de la concentration, et idées suicidaires.

Références

ADLER A., BUSH A., BARG F. K., WEISSINGER G., BECK A. T. & BROWN G. K. (2016). A mixed methods approach to identify cognitive warning signs for suicide attempts. *Archives of Suicide Research*, **20**(4), 528–538.

BARBOSA L. & FENG J. (2010). Robust sentiment detection on Twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, COLING 2010, p. 36–44, Beijing, China : Association for Computational Linguistics.

BRADLEY M. M. & LANG P. J. (1999). *Affective norms for English words (ANEW) : Stimuli, instruction manual, and affective ratings*. Rapport interne, Center for Research in Psychophysiology, University of Florida, Gainesville, Florida.

BURNAP P., COLOMBO W. & SCOURFIELD J. (2015). Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, HT '15, p. 75–84, New York, NY, USA : ACM.

COLOMBO G. B., BURNAP P., HODOROG A. & SCOURFIELD J. (2016). Analysing the connectivity and communication of suicidal users on twitter. *Computer Communications*, **73, Part B**, 291 – 300. Online Social Networks.

DE CHOUDHURY M., COUNTS S. & HORVITZ E. (2013). Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI'13, p. 3267–3276, New York, NY, USA : ACM.

GUNN J. F. & LESTER D. (2012). Twitter postings and suicide : An analysis of the postings of a fatal suicide in the 24 hours prior to death. *Suicidologi*, **17**(3), 28–30.

HOFFMAN M. D., BLEI D. M. & BACH F. (2010). Online learning for latent dirichlet allocation. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, NIPS'10, p. 856–864, USA : Curran Associates Inc.

KARMEN C., HSIUNG R. C. & WETTER T. (2015). Screening internet forum participants for depression symptoms by assembling and enhancing multiple NLP methods. *Computer Methods and Programs in Biomedicine*, **120**(1), 27–36.

KIM S., LI F., LEBANON G. & ESSA I. A. (2013). Beyond sentiment : The manifold of human emotions. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, p. 360–369.

LAFFERTY J. D., MCCALLUM A. & PEREIRA F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, p. 282–289, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

LARSEN R. J. & DIENER E. (1992). Promises and problems with the circumplex model of emotion. *Review of Personality and Social Psychology*, **13**(13), 25–59.

LEHRMAN M. T., ALM C. O. & PROAÑO R. A. (2012). Detecting distressed and non-distressed affect states in short forum texts. In *Proceedings of the 2012 Workshop on Language in Social Media*, LSM 2012, p. 9–18, Montreal, Canada : Association for Computational Linguistics.

MAIGROT C., BRINGAY S. & AZÉ J. (2016). Concept drift vs suicide : comment l'un peut prévenir l'autre ? In *16ème Journées Francophones Extraction et Gestion des Connaissances, EGC 2016*, volume E-30, p. 219–230.

MORENO M., JELENCHICK L., EGAN K., COX E., YOUNG H., GANNON K. & BECKER T. (2011). Feeling bad on Facebook : depression disclosures by college students on a social networking site. *Depression and Anxiety*, **28**(6), 447–455.

O'DEA B., WAN S., BATTERHAM P. J., CALEAR A. L., PARIS C. & CHRISTENSEN H. (2015). Detecting suicidality on twitter. *Internet Interventions*, **2**(2), 183 – 188.

PESTIAN J. P., MATYKIEWICZ P. & GRUPP-PHELAN J. (2008). Using natural language processing to classify suicide notes. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '08, p. 96–97, Stroudsburg, PA, USA : ACL.

POULIN C., SHINER B., THOMPSON P., VEPSTAS L., YOUNG-XU Y., GOERTZEL B., WATTS B., FLASHMAN L. & MCALLISTER T. (2014). Predicting the risk of suicide by analyzing the text of clinical notes. *PLoS ONE*, **9**(1).

SPASIC I., BURNAP P., GREENWOOD M. & ARRIBAS-AYLLON M. (2012). A Naïve Bayes Approach to Classifying Topics in Suicide Notes. *Biomedical Informatics Insights*, **5**(1), 87–97.

SUEKI H. (2015). The association of suicide-related twitter use with suicidal behaviour : A cross-sectional study of young internet users in japan. *Journal of Affective Disorders*, **170**, 155 – 160.

SUTTON C. & MCCALLUM A. (2012). An introduction to conditional random fields. *Found. Trends Mach. Learn.*, **4**(4), 267–373.

YIK M., RUSSELL J. A. & STEIGER J. H. (2011). A 12-point circumplex structure of core affect. *Emotion*, **11**(4), 705–731.

Une plate-forme visuelle pour une information comparative sur les nouveaux médicaments⋆

Jean-Baptiste Lamy[1], Adrien Ugon[1], Catherine Duclos[1], Alain Venot[1], Madeleine Favre[2], Hélène Berthelot[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@univ-paris13.fr, adrien.ugon@lip6.fr, catherine.duclos@avc.aphp.fr, alain.venot@univ-paris13.fr, helene.berthelot@orange.fr

[2] Dept. de médecine générale, Université Paris Descartes, Société de Formation Thérapeutique du Généraliste (SFTG), Paris
mfavre89@gmail.com

Résumé : Lorsqu'un nouveau médicament arrive sur le marché, les médecins doivent décider s'ils le considéreront pour leurs prescriptions futures ou non. La bonne décision dépend des propriétés du nouveau médicament, des médicaments déjà existants dans la même indication, mais aussi de la pratique du médecin et de sa patientèle. Elle est difficile à prendre, car une part importante de l'information sur les nouveaux médicaments provient de l'industrie pharmaceutique et n'est donc pas indépendante. Nous proposons ici un système visuel d'aide à la décision pour aider le médecin dans ce contexte, à partir d'une information indépendante, objective et fiable.

Le système présenté repose d'une part sur une ontologie comparative du médicament, qui permet la comparaison des propriétés du nouveau médicament avec celles des médicaments existants similaires, et d'autre part sur des techniques de visualisation de connaissance permettant de comparer en un coup d'oeil les propriétés de plusieurs médicaments. Nous présentons ici un prototype incluant 4 nouveaux médicaments et 22 médicaments comparateurs. L'évaluation de ce prototype par un groupe de 22 médecins généralistes montre que ce système visuel permet aux médecins de se forger leur propre opinion sur les nouveaux médicaments, et de changer d'avis sur certains médicaments.

Mots-clés : Ontologies, Visualisation des connaissance, Médicament

1 Introduction

Lorsqu'un nouveau médicament arrive sur le marché, les médecins doivent décider s'ils le prendront en considération pour leurs prescriptions futures ou non. La décision doit prendre en compte les propriétés du nouveau médicament, celles des autres médicaments déjà existants dans la même indication, mais aussi la pratique du médecin et sa patientèle. Par exemple, un nouveau médicament ayant pour effet indésirable fréquent des diarrhées peut être problématique pour un médecin qui traite beaucoup de jeunes enfants (les diarrhées pouvant être fatales dans ce cas), tandis que pour des médecins ayant une patientèle adulte, ce problème sera jugé moins grave.

L'information [1] disponible sur les nouveaux médicaments est souvent peu exploitable : il s'agit principalement d'information en provenance de l'industrie pharmaceutique, qui peut donc être biaisée, ou d'avis d'experts (par exemple dans des journaux comme Prescrire) qui

⋆. Ce travail a été financé par l'ANSM (Agence Nationale de Sécurité du Médicament et des produits de santé) au travers du projet de recherche VIIIP (AAP-2012-013).

1. L'usage veut que l'on parle d'*information* sur le médicament, bien qu'il s'agisse en réalité de *connaissance*, au sens de F Nake (Nake F, 2001), c'est-à-dire une information qui est réutilisable dans un autre contexte (par exemple pour différents patients).

ne prennent pas en compte la pratique du médecin et sa patientèle. Une possibilité serait de permettre au médecin de comparer systématiquement les propriétés du nouveau médicament (telles que les contre-indications ou les effets indésirables décrits dans les documents officiels : Résumés des Caractéristiques Produit, RCP) avec celles des médicaments déjà existants dans la même indication. Le médecin pourrait ensuite se faire sa propre opinion en tenant compte de sa situation particulière. En pratique, cela est quasi impossible à cause du nombre important de propriétés et de l'hétérogénéité dans la description d'une même propriété. Cette hétérogénéité se rencontre dans les textes mais également dans les bases de données codées sur le médicament. En effet, ces bases ont été conçues pour la consultation des propriétés d'un médicament, mais pas pour la comparaison de plusieurs médicaments. Ces bases de données ont une dimension sémantique très faible.

Dans cet article, nous proposons une plate-forme visuelle d'aide à la décision pour aider le médecin à se faire une opinion sur les nouveaux médicaments, en situation de formation. Ce système s'appuie d'une part sur une ontologie comparative des médicaments, qui permet de décrire les propriétés des médicaments et de les comparer, et d'autre part sur des méthodes de visualisation des connaissances pour rendre possible la comparaison visuelle des nombreuses propriétés d'un petit groupe de médicaments. Nous présentons un prototype de la plate-forme incluant 4 nouveaux médicaments et 22 médicaments comparateurs ; nous mettrons l'accent sur la comparaison des effets indésirables qui est complexe car il faut prendre en compte la nature des effets mais aussi leur fréquence et leur gravité. Nous présentons également l'évaluation du prototype par 22 médecins généralistes.

La suite de l'article est organisée de la manière suivante. La section 2 présente la construction d'une ontologie comparative des médicaments. La section 3 décrit la plate-forme visuelle que nous proposons pour l'information sur le nouveau médicament. La section 4 aborde l'évaluation de cette plate-forme par un petit groupe de médecins. Enfin, la section 5 discute l'approche d'aide à la décision visuelle que nous avons suivie, avant de conclure.

2 L'ontologie comparative des médicaments

Plusieurs « ontologies » du médicament ont été publiées, cependant aucune d'entre elles ne décrit l'ensemble des propriétés cliniques des médicaments. DrOn (*Drug Ontology*) (Hanna *et al.*, 2013) s'intéresse à l'identification des médicaments selon la nomenclature américaine RxNorm, mais ne renseigne ni contre-indication ni effet indésirable. PDO (*Prescription of Drugs Ontology*) (Ethier *et al.*, 2016) modélise seulement les lignes de prescription (par exemple « 1 comprimé de paracétamol 2 fois par jour »). OAE (*Ontology of Adverse Events*) (He *et al.*, 2014) décrit les événements indésirables qui ont eu lieu chez un patient donné (*adverse events*) mais pas les effets indésirables potentiels des médicaments (*adverse effects*). DID (*Drug-Indication Database*) (Sharp ME, 2017) s'intéresse aux indications, mais ne propose pas un modèle détaillé de celles-ci et se limite à des couples (médicament, pathologie traitée), l'ontologie étant décrite comme une « base de données ». OntoADR (Souvignet *et al.*, 2016) propose un modèle ontologique de la terminologie MedDRA (*Medical Dictionary for Regulatory Activities*) utilisée pour les effets indésirables, mais se limite à la nature des effets et ne les relie pas au médicament.

De plus les experts avec lesquels nous avons travaillé maîtrisaient mal l'anglais. Nous avons donc construit une ontologie des médicaments centrée sur les propriétés cliniques et cherchant

à rendre comparables ces propriétés malgré leurs descriptions hétérogènes. Cette ontologie est « comparative » dans le sens où son objectif est de permettre la comparaison des propriétés entre médicaments, ce que ne permettent pas les bases de données actuelles.

2.1 Construction de l'ontologie

Les principales catégories d'information nécessaires aux médecins pour évaluer les nouveaux médicaments ont été déterminées à partir d'études précédentes réalisées dans notre laboratoire (Iordatii *et al.*, 2013), et à partir de deux séances de focus groupes réalisées ultérieurement et ayant réuni 17 médecins généralistes au total. L'objectif de ces séances était de réaliser l'expression des besoins des généralistes en matière d'information sur les nouveaux médicaments. Les séances comportaient une discussion générale sur l'innovation pharmaceutique puis un travail personnel des médecins sur un ensemble de documents portant sur les médicaments suivants, considérés comme nouveau : Alvesco® (ciclésonide, un nouveau corticoïde pour l'asthme), Cialis® (tadalafil, une nouvelle indication de ce médicament pour l'hypertrophie bénigne de la prostate), Pyléra® (bismuth + métronidazole + tétracycline, une nouvelle thérapie pour éradiquer *H. pylori*), Jext® (adrénaline, une nouvelle forme galénique avec un stylo injecteur). Les documents incluaient des documents promotionnels issus des laboratoires, les notices patients et les RCP, l'avis de la commission de transparence, et des tableaux comparatifs incluant les prix et les effets indésirables (réalisés par HB, pharmacienne experte). Les médecins ont lu les documents, ont surligné les passages jugés importants, qui ont ensuite été analysés. Les séances ont également été enregistrées.

Ensuite, nous avons conçu une ontologie comparative des médicaments, centrée sur les nouveaux médicaments. Cette ontologie inclut les propriétés du nouveau médicament et la liste des médicaments comparables (déterminée par les experts à partir des avis de la commission de transparence, lesquels prennent en compte les indications mais aussi les niveaux de prise en charge ; par exemple on ne comparera pas le paracétamol avec la morphine, ni la metformine (antidiabétique oral) avec l'insuline injectable) et leurs propriétés. La CIM10 (Classification Internationale des Maladies, version 10) a été utilisée pour coder les contre-indications, et MedDRA (version 18) pour les effets indésirables. Le modèle obtenu a été instancié manuellement sur 15 médicaments nouveaux par les auteurs : JBL, CD, AL, HB et MF ont traité 3 médicaments chacun. Chaque jeu incluait un médicament avec un nouveau principe actif, un avec une nouvelle forme galénique ou une nouvelle voie d'administration, un avec un nouveau dosage. Le modèle a été enrichi et corrigé par les auteurs suite à cette phase de test, afin d'ajouter les éléments identifiés comme manquants. En particulier, la date de mise sur le marché a été ajoutée, et nous avons distingué les propriétés valables pour un médicament en général et celles valables seulement dans une indication donnée (certains médicaments ayant plusieurs indications). Enfin, l'ontologie a été éditée avec Protégé et formalisée en OWL 2 (*Web Ontology Language*).

2.2 Description de l'ontologie

L'ontologie appartient à la famille $\mathcal{SHOIQ}(D)$ des logiques de description. La partie générale de l'ontologie (non spécifique à un médicament donné) comprend 240 classes, 167 propriétés, 154 individus et 2071 axiomes. Sa publication n'est pas possible car elle intègre des

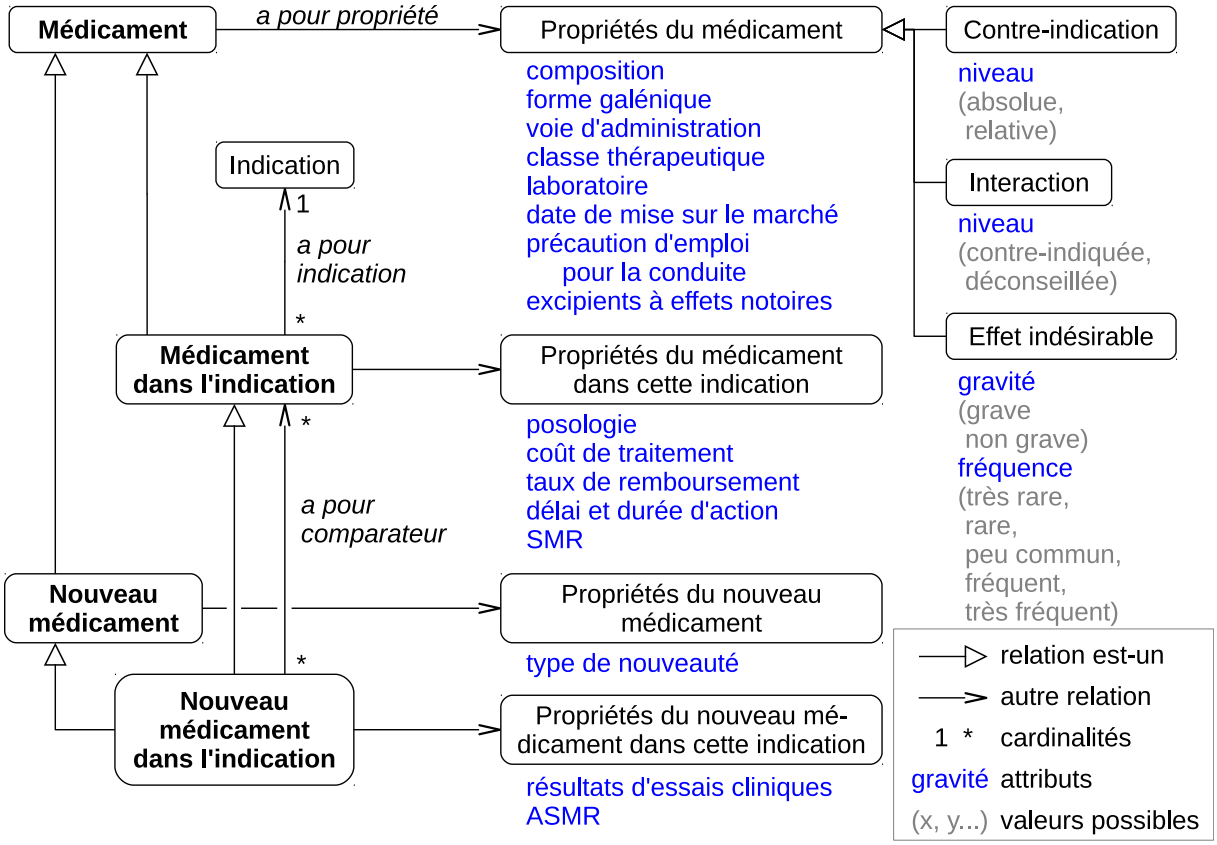


FIGURE 1 – Les principales classes de l'ontologie comparative des médicaments, les principales propriétés et le niveau où elles sont rattachées.

terminologies non redistribuables (comme MedDRA).

La Figure 1 montre les principales classes de l'ontologie et les différentes propriétés qui permettront la comparaison des médicaments. L'ontologie distingue les nouveaux médicaments (ceux-ci ne peuvent pas être déterminés automatiquement par rapport à leur date d'Autorisation de Mise sur le Marché (AMM), car la mise sur le marché effective est parfois très éloignée de la date d'AMM). Lorsqu'un médicament possède plusieurs indications, certaines propriétés sont valables pour le médicament (par exemple la composition ou les contre-indications), tandis que d'autres ne sont valables pour ce médicament que dans une indication donnée (par exemple la posologie, le coût ou la liste des comparateurs). De plus, certaines propriétés ne sont pertinentes que pour le nouveau médicament (par exemple l'ASMR, Amélioration du Service Médical Rendu, qui est donné à un temps t et pas toujours mis à jour). Par conséquent, nous avons créé quatre classes (Médicament, Médicament dans l'indication, Nouveau médicament et Nouveau médicament dans l'indication) avec des relations d'héritage, et nous avons rattaché chaque propriété à la classe appropriée. Les propriétés complexes (contre-indications, interactions et effets indésirables) sont représentées par des classes car elles ont elles-mêmes des attributs. De plus, elles peuvent être restreintes à une indication.

Pour les contre-indications, les interactions et les effets indésirables, des attributs supplémentaires sont présents : niveau de contre-indication ou d'interaction, fréquence et gravité des

effets indésirables. La liste des interactions et des effets indésirables pouvant être longue, nous l'avons limitée aux interactions de niveaux contre-indiqué et déconseillé, et aux effets indésirables graves et/ou fréquents (y compris très fréquents). Ce choix correspond aux demandes exprimées par les médecins lors des focus groupes.

Cette ontologie permet la comparaison des propriétés des médicaments. La principale difficulté rencontrée pour la comparaison est l'expression de propriétés proches à des niveaux de granularité différents, par exemple un nouveau médicament est contre-indiqué avec les maladies hémorragiques tandis qu'un de ses comparateurs est contre-indiqué avec les maladies hémorragiques *constitutionnelles* ou *acquises*, ce qui revient au même. L'ontologie permet des méthodes de raisonnements sémantiques pour résoudre ces problèmes (Lamy *et al.*, 2015). Dans l'exemple précédent, constitutionnelle et acquise réalisent une partition, que l'on peut exprimer formellement :

$Acquise \sqsubseteq Origine$
$Constitutionnelle \sqsubseteq Origine$
$Acquise \sqcap Constitutionnelle \sqsubseteq \bot$
$Origine \sqsubseteq (Acquise \sqcup Constitutionnelle)$

Il est ensuite possible de définir les trois maladies :

$Maladie \sqsubseteq (\exists aPourOrigine.Origine) \sqcap (\forall aPourOrigine.Origine)$
$MaladieHémorragique \sqsubseteq Maladie$
$MHAcquise \equiv MaladieHémorragique \sqcap \exists aPourOrigine.Acquise$
$MHConsti \equiv MaladieHémorragique \sqcap \exists aPourOrigine.Constitutionnelle$

Puis nous définissons un médicament m ayant deux contre-indications ciA et ciC :

$(ContreIndication \sqcap (\exists aPourMaladie.MHAcquise) \sqcap (\forall aPourMaladie.MHAcquise))(ciA)$
$(ContreIndication \sqcap (\exists aPourMaladie.MHConsti) \sqcap (\forall aPourMaladie.MHConsti))(ciC)$
$MHAcquise \sqsubseteq aPourMaladie^{-}.\{ciA\}$ [2]
$MHConsti \sqsubseteq aPourMaladie^{-}.\{ciC\}$
$(Médicament \sqcap (\forall aPourContreIndication.\{ciA, ciC\}))(m)$
$aPourContreIndication(m, ciA)$
$aPourContreIndication(m, ciC)$

Nous pouvons enfin définir la classe de l'ensemble des maladies contre-indiquées avec m :

$Maladie\_CI\_avec\_m \equiv Maladie \sqcap (\exists aPourMaladie^{-}.(\exists aPourContreIndication^{-}.\{m\}))$

Un raisonneur permet alors d'inférer $MaladieHémorragique \sqsubseteq Maladie\_CI\_avec\_m$ (et pas seulement $MHAcquise \sqsubseteq Maladie\_CI\_avec\_m$ et $MHConsti \sqsubseteq Maladie\_CI\_avec\_m$).

3 La plate-forme visuelle

La plate-forme contient une page pour chaque indication de chaque nouveau médicament. Des liens hypertextes permettent de naviguer entre les indications d'un même médicament. Chaque page suit la même structure générale (Figure 2) en 4 parties : (1) une zone de titre identifiant le nouveau médicament, (2) une synthèse récapitulant les propriétés du nouveau médicament seul (information non comparative), l'indication concernée et la liste des comparateurs pour cette indication, (3) une comparaison des propriétés du nouveau médicament avec

2. $aPourMaladie^{-}$ est la relation inverse de $aPourMaladie$.

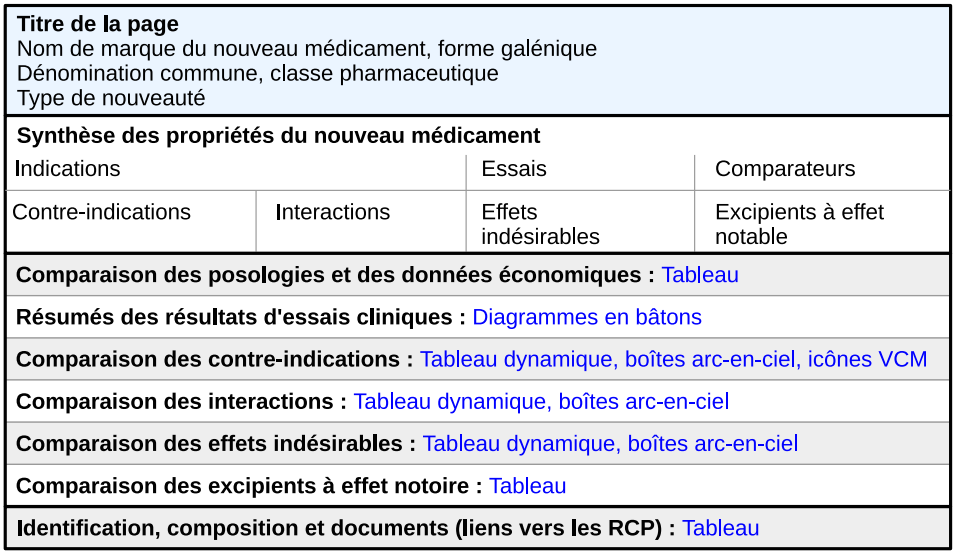


FIGURE 2 – Structure générale d'une page de la plate-forme présentant un nouveau médicament dans une de ses indications. En bleu figurent les techniques de visualisation utilisées dans chaque rubrique.

les comparateurs, divisée en 6 rubriques, et (4) un tableau identifiant chaque médicament et pointant vers les documents de référence.

La plate-forme est un site web en HTML avec des styles CSS et du JavaScript, généré par un programme Python utilisant la programmation orientée ontologie pour accéder à l'ontologie. La plate-forme avec 4 nouveaux médicaments peut être consultée à l'adresse suivante : http://www.lesfleursdunormal.fr/static/viiip_proto/html/.

En plus des techniques de visualisation simples (tableaux et diagrammes en bâton), la plate-forme combine trois techniques avancées pour comparer les propriétés des médicaments en rapport avec la sécurité (contre-indications, interactions, effets indésirables). Celles-ci posent problème car elles sont souvent très nombreuses. Lorsque plusieurs techniques sont disponibles pour une rubrique, des boutons permettent de basculer de l'une à l'autre. Les sections suivantes décrivent ces techniques avancées.

3.1 Icônes

Nous avons utilisé des icônes VCM (Visualisation des Connaissances Médicales). VCM est un langage iconique qui permet de représenter les principaux concepts médicaux, dont les maladies, par une combinaison de couleurs, formes et pictogrammes (Lamy *et al.*, 2008). Les icônes VCM ont été utilisées pour enrichir les listes de contre-indications. Elles permettent d'identifier en un coup d'oeil la présence ou l'absence de contre-indications d'un type donné (cardiaque par exemple, en recherchant visuellement le pictogramme coeur). La Figure 3 montre un exemple d'utilisation de VCM.

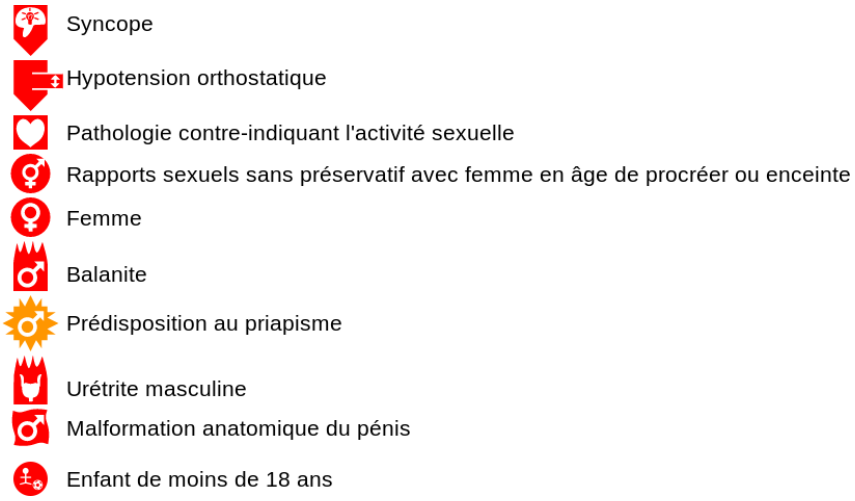


FIGURE 3 – Exemple de liste de contre-indications enrichie par des icônes VCM.

	AINS + opiacé	paracétamol + opiacé			
	Antarene Codeine	Dafalgan Codeine	Izalgi	Lamaline	Ixprim
Affections hématologiques et du système lymphatique					
Anémie hémolytique	■				
Agranulocytose	■				
Affections du système immunitaire					
Hypersensibilité médicamenteuse	■	■■	■■	■■	■■
Choc anaphylactique	■	■■	■■	■■	■■
Angioedème	■	■■	■■	■■	■■
Affections du système nerveux					
Méningite aseptique	■				
Vertige	■■■■■	■■■■	■■■■	■■■■	■■■■■
Somnolence	■■■■	■■■■	■■■■	■■■■	■■■■■
Affections cardiaques					
Infarctus du myocarde	■				
Insuffisance cardiaque	■				

FIGURE 4 – Extrait (environ 1/3) d'un tableau dynamique montrant les effets indésirables d'Antarène codéine et de quatre comparateurs. Tous les effets du nouveau médicament sont affichés, ainsi que les effets que les comparateurs partagent avec lui.

3.2 Tableaux dynamiques

Les experts médicaux utilisent généralement des tableaux pour comparer les médicaments, avec une colonne par médicament et une ligne par propriété. Ces tableaux sont faciles à comprendre mais souvent difficiles à lire à cause du grand nombre de propriétés. Nous avons amélioré ces tableaux (1) en ajoutant des symboles colorés, (2) en mettant en surbrillance les lignes pour lesquelles le nouveau médicament diffère des comparateurs et (3) en rendant le tableau interactif, afin de n'afficher qu'une partie des lignes.

Le tableau dynamique permet d'afficher : (a) les propriétés du nouveau médicament seules (celles des comparateurs ne sont affichées que si elles sont partagées avec le nouveau médicament), (b) une comparaison entre les propriétés du nouveau médicament et un comparateur choisi par l'utilisateur (en cliquant sur la colonne correspondante), cette option permet au mé-

decin de comparer le nouveau médicament avec celui qu'il a l'habitude de prescrire, (c) les propriétés partagées par la majorité des comparateurs mais absentes du nouveau médicament, ce qui permet de vérifier si le nouveau médicament peut être prescrit lorsque la plupart des médicaments existants sont contre-indiqués, et (d) toutes les propriétés de tous les médicaments. Les options (b) et (c) utilisent un raisonnement sémantique pour calculer les lignes affichées.

La Figure 4 montre un exemple de tableau dynamique affichant les effets indésirables du nouveau médicament (option (a)). Les effets graves sont en rouge, les autres en noir. Dans chaque case, la fréquence des effets est représentée par 0 à 5 petits carrés orange (0 : effet absent, 1 : effet très rare, 2 : rare, 3 : peu commun, 4 : fréquent, 5 : très fréquent).

3.3 Boîtes arc-en-ciel

La comparaison des propriétés des médicaments relève du champ de la visualisation d'ensembles non disjoints (*overlapping set visualization*). En effet, les médicaments peuvent être considérés comme des éléments, et leurs propriétés comme des ensembles regroupant les médicaments partageant une propriété commune (par exemple l'ensemble des médicaments pouvant provoquer des vomissements). Plusieurs approches ont été proposées dans ce domaine (Alsallakh *et al.*, 2014) : les diagrammes d'Euler et de Venn, la superposition de régions sur des cartes géographiques, les graphes reliant chaque élément aux ensembles auxquels il appartient, les tableaux, les données agrégées et les nuages de points. Nous avons mis au point la technique des *boîtes arc-en-ciel* (Lamy *et al.*, 2016a).

Dans les boîtes arc-en-ciel, les éléments (médicaments) sont affichés en colonne et les ensembles (propriétés) sont représentés par des boîtes rectangulaires qui recouvrent les colonnes correspondant aux éléments appartenant à l'ensemble. Les colonnes sont ordonnées à l'aide d'un algorithme heuristique pour que les éléments partageant une même propriété soient placés côte à côte. Lorsque cela n'est pas possible, un « trou » est présent dans la boîte qui comporte alors deux rectangles reliés par un trait horizontal. Les boîtes sont empilées verticalement, en plaçant les boîtes les plus larges en bas.

La Figure 5 montre un exemple de boîtes arc-en-ciel pour la comparaison des effets indésirables de cinq médicaments (un nouveau + 4 comparateurs). Le nouveau médicament est placé dans la colonne de gauche. Des couleurs ont été appliquées sur les boîtes pour indiquer la gravité et la fréquence : les effets graves ont une teinte rouge et les non graves une teinte orange, et la saturation indique la fréquence (plus la couleur est vive, plus l'effet est fréquent ; nous avons retenu 5 niveaux de fréquence pour les effets graves et 2 seulement pour les autres).

La visualisation donne une vue d'ensemble des effets indésirables et permet de répondre facilement à différentes questions que peut se poser un médecin lorsqu'il évalue un nouveau médicament : A-t-il des effets graves fréquents ? (il suffit de rechercher la couleur rouge vif dans la colonne du nouveau médicament ; ici pour Antarène codéine, il s'agit d'hématémèse et méléna). Sont-ils partagés par les médicaments similaires ? (regarder si les boîtes correspondantes se prolongent sur les comparateurs ; ici, l'hématémèse n'est pas présente chez les comparateurs et le méléna est présent chez un seul comparateur avec une fréquence plus faible). Le nouveau médicament a-t-il plus d'effets indésirables ? (regarder le nombre de boîtes dans les différentes colonnes ; ici Antarène codéine semble présenter davantage d'effets indésirables que les autres médicaments).

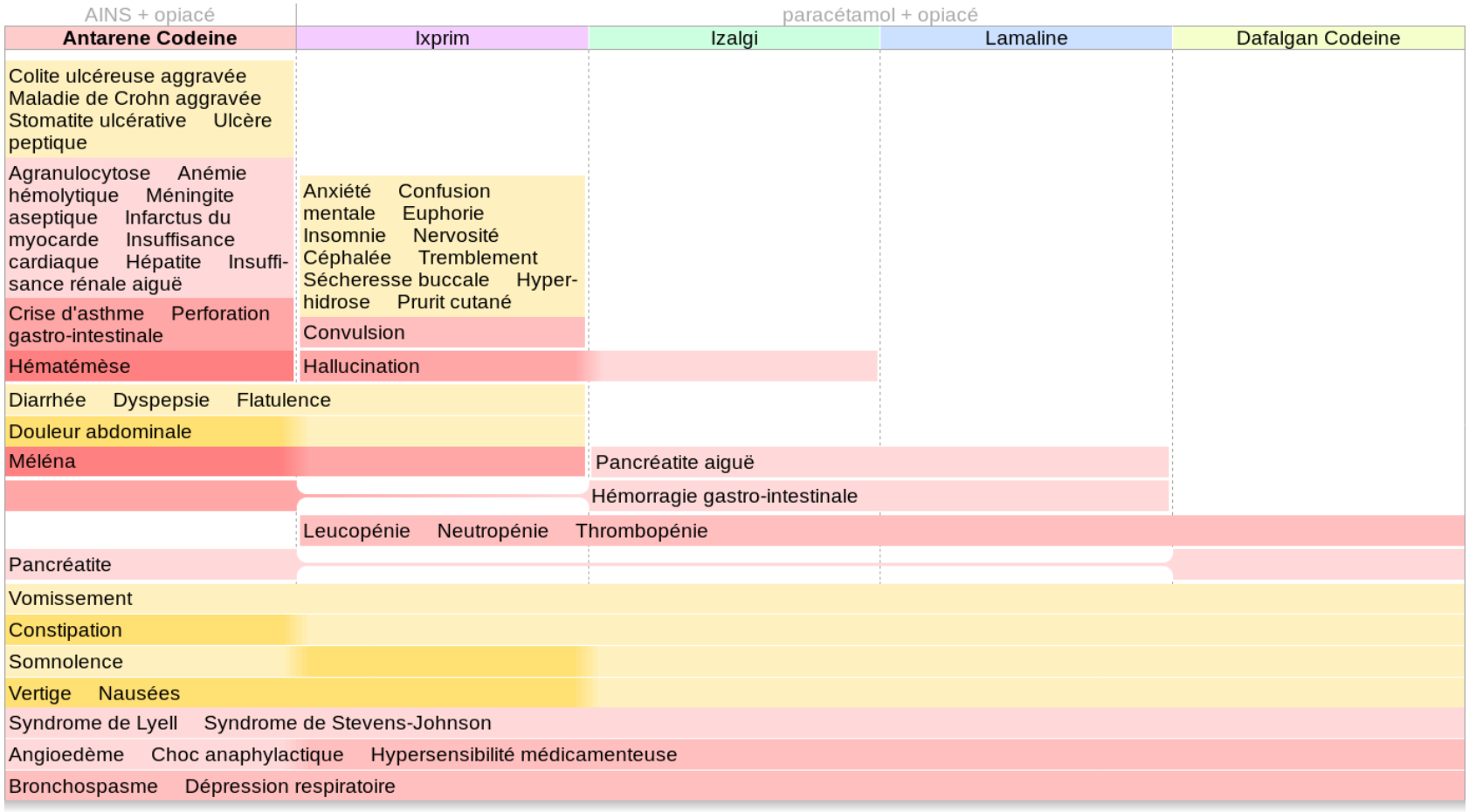


FIGURE 5 – Boîtes arc-en-ciel comparant les effets indésirables d'Antarène codéine (nouveau médicament) avec 4 comparateurs (même jeu de données que la Figure 4, mais les boîtes arc-en-ciel permettent d'afficher la totalité des effets indésirables sur un écran).

4 L'évaluation de la plate-forme

4.1 Méthodes

Pour l'évaluation, nous avons conçu un prototype comportant 4 nouveaux médicaments : Antarène codéine® (ibuprofen+codeine, pour traiter les douleurs modérées à sévères), Ciloxan® (ciprofloxacine, pour les infections auriculaires avec deux indications distinctes), Vitaros® (alprostadil, pour la dysfonction érectile) et Pyléra® (bismuth+métronidazole+tétracycline, pour l'éradication de *H. pylori*). Par rapport à la section 2.1, nous n'avons conservé qu'un seul médicament pour limiter les risques de biais (c'est-à-dire ne pas évaluer uniquement sur les médicaments que nous avons étudiés). Les informations pour ces 4 médicaments et leurs 22 comparateurs ont été extraites et codées manuellement par HB. Nous avons recruté 22 nouveaux médecins généralistes via une association de formation continue (12 hommes, 10 femmes, âge moyen : 54,6 ans).

Nous avons choisi un protocole de type « avant-après ». L'évaluation s'est déroulée en présentiel, au cours de deux séances identiques. Au début de la séance, la plate-forme était présentée rapidement aux médecins (20 minutes). Les médecins ont rempli un premier questionnaire [3] demandant, pour chacun des 4 nouveaux médicaments, s'ils manquaient d'information et s'ils étaient prêts à le prescrire en pratique. Ensuite, les médecins ont consulté la plate-forme pendant 45 minutes. Puis ils ont rempli un second questionnaire avec les mêmes questions que le premier et des questions additionnelles portant sur leur opinion sur la plate-forme. Enfin une discussion générale a été organisée.

4.2 Résultats

88 décisions ont été collectées (22 médecins $\times$ 4 nouveaux médicaments), avant et après la consultation de la plate-forme. Les décisions ont été classées selon 3 catégories : (1) le médecin manque d'information sur le nouveau médicament (ce qui, normalement, l'amène à ne pas le prescrire faute d'information), (2) le médecin a suffisamment d'information et n'est pas prêt à prescrire le nouveau médicament, (3) le médecin a suffisamment d'information et est prêt à prescrire. Avant la consultation de la plate-forme, dans 39 cas le médecin manquait d'information, dans 14 cas il n'était pas prêt à prescrire et dans 35 cas il était prêt. Après la consultation, dans 1 seul cas le médecin manquait d'information, dans 45 cas il n'était pas prêt à prescrire et dans 42 cas il était prêt. Parmi les cas où initialement le médecin manquait d'information, dans 20 cas le médecin se décide pour la non-prescription et dans 18 cas pour la prescription. De plus, dans 11 cas le médecin était prêt à prescrire et change d'avis après consultation de la plate-forme ; en revanche nous n'avons pas observé de changement d'avis dans le sens contraire.

L'ensemble des médecins ont dit que la plate-forme leur a permis de se forger une bonne idée des quatre nouveaux médicaments, et préfèrent une information comparative à une information limitée au nouveau médicament. 21 médecins recommanderaient la plate-forme à un collègue, 19 ont facilement appris à utiliser la plate-forme. Lors de la discussion générale, les médecins ont fait preuve d'un grand enthousiasme. Ils ont apprécié la neutralité de la plate-forme et trouvé que les deux outils proposés (tableaux dynamiques et boîtes arc-en-ciel) étaient complémentaires. Ils ont plus particulièrement été intéressés par la comparaison des effets indésirables.

3. Voir annexe : http://www.lesfleursdunormal.fr/static/_downloads/ic2017_annexe.pdf

5 Discussion et conclusion

Dans cet article, nous avons présenté une plate-forme visuelle pour une information comparative sur les nouveaux médicaments. Nous avons décrit l'ontologie sous-jacente à la plate-forme et les techniques de visualisation employées. Nous avons donné des résultats d'évaluation sur un petit groupe de médecins qui ont montré une bonne acceptation de leur part ainsi que la capacité de la plate-forme à faire évoluer l'avis des médecins.

Lors de l'évaluation, les médecins ont fait plusieurs propositions pour enrichir la plate-forme, avec des résultats plus détaillés sur les essais cliniques et incluant des comparaisons indirectes (nouveau médicament - comparateur *via* un placebo par exemple). Ils ont trouvé le site adapté à la formation continue (l'utilisation initialement prévue) mais aussi pour l'aide à la prescription, et souhaiteraient généraliser l'information comparative à l'ensemble des médicaments (sans se limiter aux nouveaux).

La principale limite actuelle de la plate-forme est l'extraction des informations. Nous avons tenté une extraction automatique, à partir de bases de données sur le médicament ou par des méthodes de traitement automatique de la langue naturelle (TAL) (Li *et al.*, 2013) appliquées sur les RCP (Lamy *et al.*, 2016b). Aucune méthode n'a permis d'obtenir une information permettant une vraie comparaison entre médicaments. Une méthode semi-automatique incluant TAL et vérification manuelle est sans doute la meilleure option.

Dans la littérature, la plupart des approches de visualisation d'ontologie s'appuient sur des arbres ou des graphes (Katifori & Halatsis, 2007; Dudás *et al.*, 2014). Au contraire, pour la visualisation des propriétés des médicaments, ce sont principalement des tableaux qui ont été utilisés (Iordatii *et al.*, 2015). Dans le domaine médical, la plupart des systèmes d'aide à la décision utilisent le raisonnement automatique (base de règles par exemple) pour présenter au médecin des recommandations explicites (« prescrivez ceci ou cela ») ou de déclencher des messages d'alerte lorsque le médecin a pris une mauvaise décision. Cette approche ne convient pas à notre contexte, car la bonne décision dépend de l'expérience du médecin, de ses pratiques et de ses patients, et tous ces éléments sont difficiles, voire impossibles, à modéliser et à intégrer dans un raisonnement automatique. Une autre approche est la recherche d'information (RI) : l'utilisateur formule une requête et obtient la réponse à sa question. Dans notre contexte, la RI ne convient pas car le médecin a besoin d'une *vue d'ensemble* des propriétés des médicaments et non pas d'une vue centrée sur une propriété spécifique. De plus, le médecin n'a que peu de temps à consacrer à l'expression et la saisie des requêtes. Ici, nous avons proposé une approche visuelle pour aider les médecins à prendre eux-mêmes une décision concernant les nouveaux médicaments.

En conclusion, nous avons montré que la comparaison des propriétés « brutes » des médicaments était possible avec une approche visuelle. Intégrée dans une plate-forme d'information sur les nouveaux médicaments, cette comparaison a permis à des médecins de se forger une bonne opinion sur quatre nouveaux médicaments, et de changer d'opinion le cas échéant. Ces travaux pourraient conduire à une information plus indépendante sur les médicaments, qui reposerait sur la comparaison systématique de leurs propriétés avérées plutôt que sur des avis d'experts dont l'indépendance est parfois difficile à établir et qui ne sont pas toujours adaptés à la patientèle ou à la pratique d'un médecin donné.

Références

ALSALLAKH B., MICALLEF L., AIGNER W., HAUSER H., MIKSCH S. & RODGERS P. (2014). Visualizing Sets and Set-typed Data : State-of-the-Art and Future Challenges. In *Eurographics Conference on Visualization (EuroVis)*.

DUDÁS M., ZAMAZAL O. & SVÁTEK V. (2014). Roadmapping and navigating in the ontology visualization landscape.

ETHIER J. F., TASEEN R., LAVOIE L. & BARTON A. (2016). Improving the semantics of drug prescriptions with a realist ontology. In *International Conference on Biomedical Ontology and BioCreative*.

HANNA J., JOSEPH E., BROCHHAUSEN M. & HOGAN W. R. (2013). Building a drug ontology based on RxNorm and other sources. *Journal of biomedical semantics*, **4**(1), 44.

HE Y., SARNTIVIJAI S., LIN Y., XIANG Z., GUO A., ZHANG S., JAGANNATHAN D., TOLDO L., TAO C. & SMITH B. (2014). OAE : The Ontology of Adverse Events. *Journal of biomedical semantics*, **5**, 29.

IORDATII M., VENOT A. & DUCLOS C. (2013). Designing concept maps for a precise and objective description of pharmaceutical innovations. *BMC medical informatics and decision making*, **13**, 10.

IORDATII M., VENOT A. & DUCLOS C. (2015). Design and evaluation of a software for the objective and easy-to-read presentation of new drug properties to physicians. *BMC medical informatics and decision making*, **15**, 42.

KATIFORI A. & HALATSIS C. (2007). Ontology visualization methods - A survey. *ACM Computing Surveys*, **39**(4), 10.

LAMY J. B., BERTHELOT H. & FAVRE M. (2016a). Rainbow boxes : a technique for visualizing overlapping sets and an application to the comparison of drugs properties. In *20th International Conference Information Visualisation*, volume 253-260, Lisboa, Portugal.

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., UGON A. & BERTHELOT H. (2016b). Automatic extraction of drug adverse effects from product characteristics (SPCs) : A text versus table comparison. *Stud Health Technol Inform*, **228**, 339–343.

LAMY J. B., UGON A., FAVRE M., DUCLOS C., VENOT A. & BERTHELOT H. (2015). Comparaison et visualisation des contre-indications des médicaments. In *Actes du 3ème Symposium Ingénierie de l'Information Médicale (SIIM)*.

LI Q., DELEGER L., LINGREN T., ZHAI H., KAISER M., STOUTENBOROUGH L., JEGGA A. G., COHEN K. B. & SOLTI I. (2013). Mining FDA drug labels for medical conditions. *BMC medical informatics and decision making*, **13**, 53.

NAKE F (2001). Data, Information, and Knowledge. In LIU, K. AND CLARKE, R.J. AND ANDERSEN, P.B. AND STAMPER, R.K., Ed., *Organizational Semiotics : Evolving a Science of Information Systems*, volume 41-50, Montréal, Québec, Canada : Kluwer.

SHARP ME (2017). Toward a comprehensive drug ontology : extraction of drug-indication relations from diverse information sources. *Journal of biomedical semantics*, **8**(1), 2.

SOUVIGNET J., DECLERCK G., ASFARI H., JAULENT M. C. & BOUSQUET C. (2016). OntoADR a semantic resource describing adverse drug reactions to support searching, coding, and information retrieval. *J Biomed Inform*, **63**, 100–107.

Approche numérique pour l'invalidation de liens d'identité (owl:SameAs)

Dimitrios Christaras Papageorgiou, Nathalie Pernelle, Fatiha Saïs

[1] Laboratoire de Recherche en Informatique, Université Paris Sud, Orsay, France
dimitrios.christaras-papageorgiou@u-psud.fr

[2] nathalie.pernelle@lri.fr

[3] Fatiha.Sais@lri.fr

Résumé : Au cours des dernières années, grâce à la standardisation des technologies Web sémantique, nous connaissons une production de données sans précédent, publiées en ligne sous forme de données liées. Dans ce contexte, lorsqu'un lien typé est déclaré entre deux ressources distinctes faisant référence à la même entité du monde réel, l'utilisation du owl :sameAs est généralement prédominant. Toutefois, des travaux récents dans la communauté des données liées ont montré des problèmes dans l'utilisation des liens owl :sameAs. Les problèmes surviennent à la fois dans les cas où ces liens sont erronés ou lorsqu'ils traduisent un lien moins strict que la sémantique des liens owl:sameAs définie dans OWL. Dans ce travail, nous présentons une méthode d'invalidation numérique de liens d'identité s'appuyant sur un calcul de similarité et sur des axiomes de l'ontologie pour détecter des liens d'identité invalides. Nous présentons nos premiers résultats expérimentaux, obtenus sur un jeu de données de la compétition internationale OAEI.

Mots-clés : Liage de données, Liens d'identité, Invalidation de liens, Ontologies, Qualité des liens et des données, Web Sémantique.

1 Introduction

Aujourd'hui, nous connaissons une production sans précédent de données, publiées sur le Linked Open Data (LOD), appelé aussi Web des Données. Cela a conduit à la création d'un espace de données global contenant des milliards d'assertions (Bizer *et al.* (2009)). Dans cet espace global, les fournisseurs de données établissent des liens RDF entre des ressources, représentées par des URIs, qui se réfèrent à la même entité du monde réel. Ainsi, on enrichit les connaissances relatives à une ressource spécifique et, par conséquent, la connaissance globale dans le Web des Données. La plupart des liens RDF reliant des ressources provenant de sources de données différentes sont des liens d'identité RDF représentés par le prédicat *owl*:*sameAs* dont la sémantique est défini dans Dean *et al.* (2004). Malheureusement, de nombreux liens d'identité existants ne reflètent pas une véritable identité et leur présence peut conduire à inférer des informations erronées et même contradictoires. Halpin *et al.* (2010) a ainsi montré que 37% de 250 liens d'identité choisis aléatoirement ont été déclarés comme erronés par cinq juges. De même, Jaffri *et al.* (2008) ont évalué la qualité du résultat d'une méthode de liage appliquée aux données de DBLP et DBpedia en choisissant arbitrairement 49 noms parmi les 491 796 auteurs disponibles dans DBLP de 2006. Ils ont montré que 92 % de ces 49 auteurs avaient alors des publications incorrectement affiliées. Compte tenu du volume important des données, les liens d'identité sont en effet souvent générés par des méthodes automatiques, avec des taux de précision inférieurs à 100 %. Les erreurs peuvent être dûes à la variation du niveau de qualité des données (i.e., complétude, correction, fraîcheur, fiabilité, etc.) entre les différentes sources de données liées, où à la difficulté de définir des règles de liage valides quelles que soient les

sources et les entités considérées. Ce problème montre la nécessité de définir des approches permettant de s'assurer de la qualité des liens. Certaines approches ont effectué une validation de type crowdsourcing pour évaluer et corriger les liens (Halpin *et al.* (2010)). D'autres approches utilisent la sémantique du lien d'identité, les axiomes de l'ontologies ou encore des hypothèses sur les données pour détecter autmatiquement qu'un lien conduit à une base de connaissance incohérente (de Melo (2013); Papaleo *et al.* (2014)). de Melo (2013) propose ainsi d'utiliser l'hypothèse du nom unique et la transitivité des liens d'identités pour détecter des inconsistances et utilise un algorithme de relaxation de contraintes pour supprimer des liens erronés. Papaleo *et al.* (2014) utilisent certains axiomes de l'ontologie (fonctionnalité) et la complétude locale de certaines propriétés pour invalider certains liens. D'autres travaux utilisent ces axiomes dans un cadre argumentatif pour générer des explications qui peuvent aider les experts à corriger les faits erronés (Arioua *et al.* (2016)). Les résultats d'approches telles que (Papaleo *et al.* (2014); de Melo (2013)) montrent que la précision des outils de liage peut réellement être améliorée quand elles sont utilisées pour filtrer les résultats. Cependant, dans un cadre purement logique, un fait erroné suffit à rendre la base de connaissances incohérente et ne permet pas de distinguer les cas où différents faits peuvent laisser penser que le lien est erroné de ceux dus à la présence d'un seul fait qui apparaît comme contradictoire.

Dans cet article, nous proposons une approche d'invalidation numérique de liens d'identité fondée sur des axiomes de fonctionnalité et sur les hypothèses de complétude-locale qui peuvent être déclarés pour certaines propriétés. Ce travail est une extension de Papaleo *et al.* (2014), dans lequel les axiomes sont utilisés dans un cadre numérique où différentes mesures d'agrégation simples peuvent être utilisées. Une première expérimentation a été menée sur des données de la compétition OAEI.

2 Approche numérique pour l'invalidation de liens d'identité

Le problème de détection de liens d'identité invalides se pose lorsque l'on souhaite vérifier si un lien d'identité *owl*:*sameAs* est valide entre deux ressources x et y dans un graphe RDF décrivant x et y. Plus précisément, nous souhaitons associer à un lien *owl*:*sameAs* un degré de confiance basé sur la similarité des descriptions des deux ressources. Notre approche exploite un graphe contextuel à profondeur n dont le contenu est délimité par un ensemble de propriétés P déclarées dans l'ontologie comme fonctionnelles, inverses-fonctionnelles ou locales-complètes. Plus précisément, la notion de graphe contextuel à profondeur n peut être défini comme suit (Papaleo *et al.* (2014)) :

Définition (**Graphe RDF**). Soit un ensemble U d'URIs, un ensemble B de nœuds blancs et un ensemble L de littéraux, un triplet RDF $< s, p, o >$ est tel que le sujet $s \in (U \cup B)$, le prédicat $p \in U$ et l'objet $o \in (U \cup B \cup L)$. Un graphe RDF G est une collection de triplets RDF.

Définition (**Chemin de propriétés de longueur** n) Soient G un graphe RDF, s un nœud dans G, et étant donné P un ensemble de propriétés défini pour G, un chemin de propriété $w_{n,s,P}$ de longueur n est une séquence alternant des nœuds et des propriétés, initiée par le nœud représentant la ressource s, $\{v_0 \equiv s, p_0, v_1, p_1, v_2, \ldots, v_{n-1}, p_{n-1}, v_n\}$, telle que : $v_0, \ldots, v_{n-1}$ sont des ressources dans G, $\forall i = 0, \ldots, n-1$ $v_i \in U$, v_n est un littéral dans G, $v_n \in L$, chaque triplet $\{v_i, p_i, v_{i+1}\}$ est une séquence dans un graphe RDF G telle que $p_i \in P$, toutes les ressources d'un chemin sont deux-à-deux distinctes.

Ce chemin peut être vu comme une collection d'assertions sélectionnées pour une ressource de départ s et un ensemble de propriétés P.

Définition (**Graphe Contextuel $G_{\{m,s,P\}}$ à profondeur** m). Soient G un graphe RDF et $s \in U$, un nœud de G, un nombre entier m et un ensemble P de propriétés défini pour G, un graphe contextuel $G_{\{m,s,P\}}$ à profondeur m pour une ressource s est un sous-graphe de G tel que chaque nœud $v_i \in G_{\{m,s,P\}}$ appartient à un chemin de propriétés de longueur n, avec $n \leq m$.

Un graphe contextuel à degré m pour une ressource s peut être considéré comme un sous-ensemble des informations pertinentes pour s, délimité par l'ensemble de propriétés P.

Définition (**Similarité contextuelle entre deux ressources** $CSim_{\{P,m\}}(x,y)$). Soient $G_{\{m,x,P\}}$ et $G'\{m,y,P\}$ deux graphes contextuels à profondeur m pour x et y, avec $P = DP \cup OP$ le sous-ensemble des propriétés de type *owl*:*DatatypeProperty* (DP) et de type *owl*:*ObjectProperty* (OP) délimitant le contexte dans les deux graphes G et G'. La similarité contextuelle pour les deux ressources x et y peut être définie comme suit :

$$CSim_{\{P,m\}}(x,y) =$$

$$F\Big(\bigcup_{\forall p_i \in DP} Sim(p_i.value(x), p_i.value(y))) \bigcup_{\forall p_j \in OP} CSim_{\{P,m\}}(p_j.value(x), p_j.value(y))\Big)$$

où :

— $p_i.value(x)$ permet d'obtenir la valeur ou les valeurs (en cas de propriétés multi-valuées) d'une propriété p_i de $G_{\{m,x,P\}}$,

— $Sim(v_x, v_y)$ est une fonction qui calcule un score de similarité dans $[0..1]$ entre v_x et v_y. Il s'agit soit de mesures de similarité élémentaires (e.g. Jacard, Jaro, Lenvenstein), soit de mesures de similarité entre ensembles de valeurs dans le cas de propriétés multi-valuées,

— F est une fonction d'aggrégation telle que la moyenne ou le minimum.

Définition du problème de détection de liens d'identité invalides : Étant donné un graphe RDF G, deux ressources x et y du graphe G, le triplet $< x, owl{:}sameAs, y >$ appartenant à G, un ensemble de propriétés P de G, un nombre entier m représentant la profondeur du graphe contextuel, deux graphes contextuels $G_{\{m,x,P\}}$ pour x et $G'_{\{m,y,P\}}$ pour y, un seuil de similarité $T \in [0..1]$, le problème d'invalidation numérique revient à déterminer si pour un couple de ressources x et y, on a : $CSim_{\{P,m\}}(x,y) \leq T$.

Comme cela a déjà été montré dans l'approche logique d'invalidation de liens d'identité Papaleo *et al.* (2014) le choix du sous-ensemble de propriétés à considérer peut être guidé par certains axiomes de l'ontologie : les propriétés fonctionnelles et inverses fonctionnelles et les propriétés locales complètes. En effet, quand une propriété p_1 est fonctionnelle, sa sémantique logique peut être exprimée par : $p_1(r,v) \wedge p_1(r,v') \Rightarrow v = v'$. Aussi la présence de valeurs différentes peut participer à l'invalidation d'un lien. De plus, si l'hypothèse du monde clos est en général inappropriée pour le Web sémantique (Heflin & Muñoz-avila (2002)), dans certains domaines et contextes spécifiques, la complétude locale de certaines propriétés peut être garantie (Wagner (2003)). Un bon exemple de propriété locale complète peut être la liste des auteurs d'un article dans un jeu de données tel que DBLP. Pour utiliser le fait qu'une propriété p est

locale-complète dans le calcul de similarité, nous vérifions que les ensembles de valeurs sont identiques. Aussi, nous vérifions que les deux ensembles de valeurs sont de même taille (si ce n'est pas le cas le score de similarité est mis à zéro), et si c'est le cas nous agrégeons les scores de similarité des paires de valeurs mises en correspondances en utilisant la fonction d'agrégation F qui a été sélectionnée.

En Figure 1, nous montrons un exemple de graphes contextuels extraits pour chercher à invalider un lien d'identité entre deux livres $b1$ et $b2$. La profondeur $m = 2$ et l'ensemble de propriétés P a été défini comme *{titre, annéeEd, auteur, nom}*. La propriété *ref*, étant non fonctionnelle et non locale complète, n'est pas prise en compte dans les graphe contextuels. La propriété *auteur* est déclarée comme locale complète. Considérons un score de similarité de 1 pour les valeurs de propriétés *titre* et *éditeur*, un score de 0 pour *annéeEd* et un score de 0.5 pour la propriété *auteur* en utilisant la mesure de similarité Jaccard. La fonction $CSim(b1, b2)$ avec $F = Moyenne$ donnerait un degré de confiance de 0.5 (0 avec $F = Minimum$).

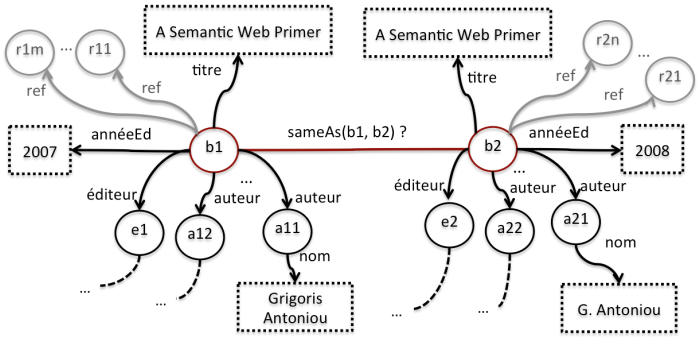


FIGURE 1 – Graphes contextuels extraits pour les ressources $b1$ et $b2$ (m=2)

3 Expérimentations

Nous avons évalué notre approche d'invalidation numérique de liens *owl :sameAs* sur les données du track PR de la compétition internationale OAEI [1] (Ontology Alignment Evaluation Initiative) 2010. Le benchmark *Person-Restaurants* a été conçu dans le but de mettre en compétition différents outils de liage de données. Les jeux de données *Person1* et *Person2* contiennent des données réelles et modifiées décrivant des personnes (SSN, nom, prénom, tél, et adresses décrites par des rues et villes), issues du projet Febrl [2]. Les données sont dupliquées de façon à ce que chaque description de *Person1* (resp. *Person2*) ait un doublon avec une modification au maximum (resp. trois modifications) par propriété. Le troisième jeu de données (*Restaurant*) a été créé en utilisant les descriptions de restaurants provenant de deux sources de données différentes. Les restaurants sont décrits par leur nom, adresse (rue, quartier et ville), téléphone et catégorie de restaurant (décrite par un nom de catégorie). Dans les trois jeux de données, le nombre d'instances varie entre 500 et 600. Pour chaque jeu de données, un goldstandard représentant le résultat de référence (i.e., un ensemble de 112 liens corrects) a été fourni. Pour tester

1. `http://oaei.ontologymatching.org/2010/`
2. `http://sourceforge.net/projects/febrl/`

notre approche, nous avons remplacé aléatoirement 50% des liens corrects par des liens erronés. Sur les trois jeux de données nous avons fait varier les mesures de similarité élémentaires (e.g.Jaro-Winkler, Jaccard), la fonction d'agrégation des scores de similarité ainsi que le seuil de similarité en dessous duquel un lien d'identité sera considéré comme invalide. La figure 2 montre les résultats en terme de rappel, précision et F-mesure obtenus sur les jeux (*Person1* et *Person2*). Nous avons comparé les résultats obtenus en utilisant la fonction moyenne (courbes en violet) et en utilisant la fonction minimum (courbes en jaunes) pour l'agrégation des scores de similarité. L'utilisation de la moyenne permet d'obtenir de bien meilleurs résultats que ceux obtenus avec un minimum, en terme de précision (et donc de F-Mesure). En effet, lorsque l'on utilise la fonction minimum, il suffit d'avoir une propriété pour laquelle le score de similarité est en dessous du seuil pour que cette similarité soit répercutée sur la similarité globale et conduise ainsi à une décision d'invalidation de lien (raisonnement analogue à celui de l'approche logique définie dans Papaleo *et al.* (2014)).

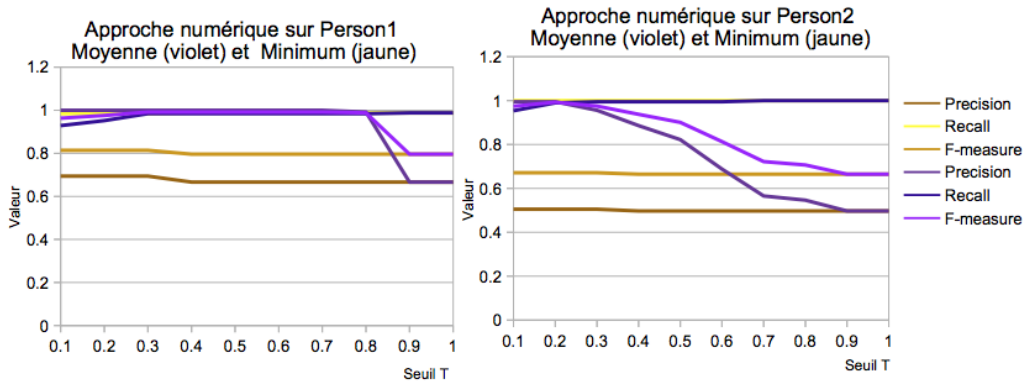


FIGURE 2 – Résultats (Précision, Rappel et F-mesure) de l 'approche numérique sur *Person1*

Les résultats obtenus sur *Restaurant*, sont similaires et montrent un gain maximum de F-Mesure de 20% quand la moyenne est utilisée. Les jeux sont très peu volumineux mais l'on peut noter que le temps d'exécution varie de 66 à 91 secondes pour les trois jeux de données (processeur Intel(R) Core(TM) i7-3630QM CPU@2.40GHz, mémoire RAM de 8Go).

Comparaison des résultats de l'approche numérique et de l'approche logique : Nous avons comparé nos résultats avec ceux obtenus par l'approche logique développée dans (Papaleo *et al.* (2014)). La Table 1 présente les résultats obtenus pour les deux approches sur les trois jeux de données *Person1*, *Person2* et *Restaurant*. Les résultats de l'approche numérique sont ceux obtenus au meilleur seuil de similarité et en utilisant la moyenne comme fonction d'agrégation des scores de similarité élémentaires. Sur les trois jeux de données les résultats de l'approche numérique en terme de F-mesure et précision sont meilleurs que les résultats de l'approche logique. En effet, nous obtenons un gain moyen de 23% de F-mesure en utilisant l'approche numérique, grâce à une augmentation très significative de la précision tout en ayant un résultat comparable en terme de rappel. Il suffit en effet d'avoir une seule propriété ayant des valeurs différentes pour que l'approche logique invalide le lien d'identité.

/	Approche logique (Papaleo *et al.* (2014))			Approche numérique			
Datasets	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure	Seuil
Person1	0.69	0.98	**0.81**	1.0	0.98	**0.99**	0.3
Person2	0.5	**1.0**	**0.67**	0.994	**0.989**	**0.99**	0.2
Restaurant	0.63	0.97	**0.77**	0.97	1.0	**0.98**	0.4

TABLE 1 – Comparaison entre l'approche logique Papaleo *et al.* (2014) et l'approche numérique

4 Conclusion

Nous avons présenté dans cet article une approche d'invalidation numérique de liens d'identité fondée sur le calcul d'un degré de confiance. Ce dernier exploite des sous-graphes RDF contextuels construits en prenant en compte des axiomes de (inverse) fonctionnalité des propriétés ainsi que des connaissances sur la complétude-locale de certaines propriétés. Les premières expérimentations ont montré la pertinence du choix d'une fonction d'agrégation simple comme la moyenne et le gain très significatif (de l'ordre de 23%) des résultats d'une telle approche par rapport à une approche purement logique (Papaleo *et al.* (2014)). Nous envisageons d'évaluer notre approche sur des jeux de données plus conséquents et de domaine différents. Nous souhaitons également combiner cette approche avec des approches de liage de données utilisant des règles efficaces de liage mais avec des exceptions.

Références

ARIOUA A., CROITORU M., PAPALEO L., PERNELLE N. & ROCHER S. (2016). On the explanation of sameas statements using argumentation. In *Scalable Uncertainty Management - 10th International Conference, SUM 2016, Nice, France, September 21-23, 2016, Proceedings*, p. 51–66.

BIZER C., HEATH T. & BERNERS-LEE T. (2009). Linked data - the story so far. *International Journal Semantic Web Information Systems*, **5**(3), 1–22.

DE MELO G. (2013). Not quite the same : Identity constraints for the web of linked data. In M. DESJARDINS & M. L. LITTMAN, Eds., *AAAI* : AAAI Press.

DEAN M., SCHREIBER G., BECHHOFER S., VAN HARMELEN F., HENDLER J., HORROCKS I., MCGUINNESS D. L., PATEL-SCHNEIDER P. F. & STEIN L. A. (2004). Owl web ontology language reference. *W3C Recommendation February*, **10**.

HALPIN H., HAYES P. J., MCCUSKER J. P., MCGUINNESS D. L. & THOMPSON H. S. (2010). When owl :sameas isn't the same : An analysis of identity in linked data. In *The Semantic Web – ISWC 2010 : 9th International Semantic Web Conference*, p. 305–320 : Springer Berlin Heidelberg.

HEFLIN J. & MUÑOZ-AVILA H. (2002). LCW-based agent planning for the semantic web. In *Ontologies and the Semantic Web Workshop*, p. 63–70 : AAAI Press.

JAFFRI A., GLASER H. & MILLARD I. (2008). Uri disambiguation in the context of linked data. In *Linked Data on the Web - LDOW*, volume 369 of *CEUR Workshop Proceedings* : CEUR-WS.org.

PAPALEO L., PERNELLE N., SAÏS F. & DUMONT C. (2014). Logical detection of invalid sameas statements in RDF data. In *Knowledge Engineering and Knowledge Management - 19th International Conference, EKAW 2014, Linköping, Sweden, November 24-28, 2014. Proceedings*, p. 373–384.

WAGNER G. (2003). Web rules need two kinds of negation. In *Principles and Practice of Semantic Web Reasoning*, volume 2901 of *LNCS*, p. 33–50. Springer Berlin Heidelberg.

Détection de liens d'identité contextuels dans une base de connaissances

Joe Raad, Nathalie Pernelle, Fatiha Saïs

LRI, Paris Sud University, Bât. 650, F-91405 Orsay, France
firstname.lastname@lri.fr

Résumé : De nombreuses applications du Web de données exploitent des liens d'identités déclarés à l'aide du constructeur $owl:sameAs$. Cependant, différentes études ont montré qu'une utilisation abusive de ces liens peut conduire à des inférences erronées ou contradictoires. Dans ce papier nous proposons de calculer des liens d'identités contextuels qui permettent d'expliciter les contextes dans lesquels ces liens sont valides. La notion de contexte que nous proposons est représentée en se basant sur l'ontologie de domaine dans laquelle les instances sont représentées. Nous avons expérimenté cette approche dans le domaine des données scientifiques où les éléments décrivant les expériences partagent rarement un lien d'identité tel que défini par $owl:sameAs$.

Mots-clés : Liage de données, Contextes, Ontologies, Bases de connaissances, Enrichissement.

1 Introduction

Le Linked Open Data cloud est une initiative du W3C, qui définit un ensemble de bonnes pratiques pour publier et lier des données structurées sur le Web. En utilisant des technologies du Web sémantique, des applications peuvent partager, extraire, interroger ou raisonner sur les données publiées. Le Linked Open Data (LOD) a récemment pris une nouvelle dimension avec la publication de grandes quantités de données (le LOD est passé de 500 millions de triplets RDF en 2007 à 130 milliards de triplets en 2016). Ces données sont encyclopédiques telles que DBpedia, Yago ou encore Google Knowledge Vault ou concernent différents domaines d'application comme les sciences du vivant, la culture ou encore l'économie. Toutefois, si ces données se retrouvent isolées, leur utilité reste très limitée. En effet, un des points angulaires du Web des données est le fait que les données soient liées entre elles par des liens sémantiques tels que les liens d'identité *owl :sameAs* qui expriment que deux ressources différentes réfèrent à la même entité (e.g., même personne, même article, même gène).

Ce type de liens est défini dans (Patel-Schneider *et al.* (2004)) avec une sémantique très stricte qui exprime le fait que déclarer un fait *owl :sameAs* entre deux objets indique que toutes les valeurs de propriétés déclarées pour l'un peuvent aussi être déclarées pour l'autre. Ainsi, si des faits *owl :sameAs* erronés sont déclarés dans les bases de connaissances cela peut conduire à inférer des informations erronées et même contradictoires. Dans (Jaffri *et al.* (2008)), les auteurs ont évalué la qualité du résultat du liage de données obtenu entre des données DBLP et des données de DBpedia. Pour cela, ils ont mesuré la correction des nouveaux faits inférés en exploitant la sémantique logique des liens *owl*:*sameAs*. En choisissant arbitrairement 49 noms parmi les 491 796 auteurs disponibles dans DBLP 2006, ils ont montré que 92 % de ces 49 auteurs ont eu des publications affiliées incorrectement. Par ailleurs, l'absence d'autres types de liens alternatifs avec une sémantique claire, renforce l'utilisation abusive du *owl*:*sameAs*. Souvent, la relation d'identité entre objets est plus faible et dépend du contexte dans lequel on

veut utiliser l'identité. Prenons l'exemple de deux éditions distinctes du même livre. Dans le cas où l'on s'intéresse à identifier s'il s'agit de la même édition de livre alors ces deux livres seront différents. Cependant, dans un contexte où l'on cherche à savoir s'il s'agit de la même oeuvre artistique alors ces deux livres seront identiques.

Dans cet article, nous proposons une approche de détection de *liens d'identité contextuels* dans des bases de connaissances RDF [1]. Un lien d'identité contextuel exprime une relation d'identité entre deux instances, valide dans un contexte défini par rapport à une ontologie de domaine. Pour cela, nous avons défini une notion de contexte global composé d'un ensemble de contextes locaux. Nous avons développé un algorithme de détection de liens d'identité contextuels qui permet de déterminer pour chaque couple d'instances les contextes globaux les plus spécifiques dans lesquels ces instances peuvent être considérées comme étant identiques. Nous avons testé notre approche sur des données scientifiques issues d'un projet dans le domaine de l'agro-alimentaire.

Dans ce qui suit, nous présentons en section 2 un ensemble de travaux connexes. Puis, en section 3, nous fournissons les objectifs ainsi que les définitions utilisées dans notre approche. En section 4 nous présentons l'algorithme DECIDE qui calcule les liens d'identité contextuels. Enfin, nous présentons les premiers résultats d'expérimentation en section 5.

2 Travaux connexes

Les approches de découverte de liens d'identité permettent de détecter que deux descriptions se réfèrent au même objet du monde réel (e.g. même personne, même lieu, même gène). Avec l'initiative du Linked Open Data cloud (LOD) proposée par Tim-Berners Lee en 2007, un fort engouement a été constaté pour le développement d'approches de liage de données RDF, dans le domaine du Web sémantique (voir Ferrara *et al.* (2011) pour un état de l'art). Pour représenter les liens d'identités générés par des approches (semi)-automatiques, le constructeur $owl{:}sameAs$ défini dans le langage OWL2 (Patel-Schneider *et al.* (2004)) est utilisé, mais sa sémantique stricte exige que, si deux objets sont liés par un lien *owl*:*sameAs*, ces derniers doivent avoir les mêmes valeurs pour toutes les propriétés ($(owl{:}sameAs(i_1, i_2) \wedge p(i_1, v) \Rightarrow p(i_2, v))$). Certaines approches se sont focalisé sur la détection de liens erronés. Dans (de Melo (2013)), les auteurs ont exploité l'hypothèse du nom unique pour détecter des liens d'identité erronés. Dans le même esprit, le travail de Ding *et al.* (2010) propose une approche globale qui par analyse de la topologie du graphe des liens d'identité, détecte des liens d'identité invalides. Enfin, l'approche logique proposée dans (Papaleo *et al.* (2014)) permet de détecter des liens d'identité invalides en exploitant un sous-graphe RDF construit en exploitant des axiomes de fonctionnalité et de complétude locale des propriétés.

D'autres approches ont caractérisé les différentes situations où le lien *owl*:*sameAs* serait utilisé à mauvais escient (Halpin *et al.* (2010); de Melo (2013); Ding *et al.* (2010)). Halpin et al. (Halpin *et al.* (2010)) ont fait état de quatre cas où le lien *owl*:*sameAs* serait utilisé de façon inappropriée. Les auteurs citent en particulier le cas où un lien d'identité est déclaré entre deux objets dont les informations décrivent une entité dans deux contextes différents. En d'autres termes, il ne s'agit pas des mêmes informations qui sont renseignées dans tous les contextes (e.g. contexte social vs professionnel). Dans ce type de cas, il est nécessaire de pouvoir distinguer et repré-

1. https ://www.w3.org/RDF/

senter les contextes dans lesquels un lien d'identité serait valide. Il existe quelques propositions pour la représentation de liens d'identité faible tels que les prédicats SKOS (Miles & Bechhofer (2009)) *skos*:*exactMatch*, *skos*:*closeMatch*, *skos*:*broadMatch* ou encore *skos*:*narrowMatch*. Ces prédicats ne permettent néanmoins pas de répondre au problème de l'identité contextuelle. De plus, les prédicats SKOS ne peuvent être utilisés que pour des URI dont le type est un concept SKOS ce qui limite les cas d'utilisation possibles dans le LOD. Dans (Halpin *et al.* (2010)), les auteurs ont développé une ontologie de l'identité dans laquelle treize prédicats ont été hiérarchisés par la relation *rdfs*:*subPropertyOf* et caractérisés par les propriétés de réflexivité, de transitivité et de symétrie. Parmi ces prédicats, on trouve les prédicats SKOS cités précédemment mais également de nouveaux prédicats comme *id*:*claimsIndentical*, *id*:*matches* et *id*:*similar*. Le prédicat *owl*:*sameAs*, qui est réflexif, symétrique et transitif, est présenté au niveau le plus spécifique (i.e., le plus strict) et au niveau le plus général on trouve le prédicat *id*:*claimsSimilar* (réflexif, non-symétrique et non-transitif) et le prédicat *id*:*claimsRelated* (non-réflexif, non-symétrique et non-transitif). Les prédicats préfixés par le mot *claims* expriment une relation d'identité ou de similarité subjective dont la véracité dépend du contexte et/ou de l'interprétation du décideur humain. Bien que cette ontologie soit précise du point de vue structuration des relations d'identité et de similarité, elle ne permet néanmoins pas d'expliciter les contextes dans lesquels une relation d'identité serait valide. Dans le but d'utiliser des liens d'identité dans des contextes spécifiques, les auteurs de (Halpin *et al.* (2015)) ont proposé l'utilisation de graphes nommés pour représenter les contextes. Les différents contextes pertinents ne sont pas toujours faciles à expliciter, même pour un expert de domaine. Aussi, plus récemment, une approche développée par (Beek *et al.* (2016)) permet de représenter l'ensemble des contextes possibles dans lesquels un lien d'identité pourrait être valide. Un contexte associé à un lien d'identité entre deux instances i_1 et i_2 correspond à un sous-ensemble de propriétés pour lesquelles i_1 et i_2 doivent avoir les mêmes valeurs. Cependant, dans cette approche, un contexte est un ensemble non structuré de propriétés qui ne tient pas compte de l'organisation multi-classes de données RDF associées à une ontologie. Par ailleurs, les auteurs ne présentent pas d'algorithme permettant de calculer les liens d'identité contextuels.

3 Une relation d'identité contextuelle guidée par l'ontologie et des connaissances expertes

L'objectif de notre approche est de découvrir des liens d'identité valides dans un contexte qui peut être défini comme une sous-partie d'une ontologie de domaine. Dans cette section, nous présentons tout d'abord le modèle de données sur lequel s'appuie notre approche. Puis, nous définissons les notions de contextes locaux et globaux utilisés pour représenter la relation d'identité proposée. Enfin, nous donnons la définition de la relation d'identité contextuelle que l'approche que nous avons développée permet de découvrir.

3.1 Modèle de données

Notre approche de détection de liens d'identité contextuelle s'appuie sur une base de connaissances où l'ontologie est représentée en OWL [2] et les données sont représentées en RDF. Une

2. https ://www.w3.org/OWL/

base de connaissances $\mathcal{B}$ est définie par un couple $(\mathcal{O}, \mathcal{F})$ où :

— $\mathcal{O} = (\mathcal{C}, \mathcal{DP}, \mathcal{OP}, \mathcal{A})$ représente la partie conceptuelle de la base de connaissances définie par un ensemble de classes $\mathcal{C}$, un ensemble $\mathcal{DP}$ de propriétés de type *owl*:*DataTypeProperty*, un ensemble $\mathcal{OP}$ de propriétés de type *owl*:*ObjectProperty* et par un ensemble d'axiomes $\mathcal{A}$ tels que la relation de subsomption entre classes, la disjonction entre classes ou encore la fonctionnalité des propriétés.

— $\mathcal{F} = \{(s, p, o)\}$ est une collection de triplets (faits) de la forme *(sujet, propriété, objet)*, exprimant des liens entre deux instances de classe ou une instance et une valeur littérale. On notera $\mathcal{I}$ l'ensemble des instances de $\mathcal{C}$ correspondant à l'ensemble des sujets s et des objets o tels que $\exists(s, p, o) \in \mathcal{F}$.

Une ontologie OWL peut être représentée par un graphe $\mathcal{G} = (V, E)$ où l'ensemble de sommets V sont les classes et les types de données élémentaires (e.g. String, Date, Integer), et les arcs E sont les propriétés liant les classes entre elles ou liant les classes à des types de données élémentaires.

3.2 Présentation du problème de détection de liens d'identité contextuels

Le problème de détection de liens d'identité contextuels auquel nous nous intéressons peut être défini comme suit : étant donnée une base de connaissances $\mathcal{B} = (\mathcal{O}, \mathcal{F})$ et l'ensemble $\mathcal{I}^c$ des instances d'une classe cible c de l'ontologie $\mathcal{O}$, trouver pour l'ensemble des paires d'instances $(i_1, i_2) \in (\mathcal{I}^c \times \mathcal{I}^c)$ les contextes les plus spécifiques pour lesquels (i_1, i_2) sont identiques.

Par exemple, on s'intéresse en Figure 1 à l'identité des individus de la classe cible Jus pour laquelle deux instances jus1 et jus2 sont représentées. L'un des contextes dans lequel les

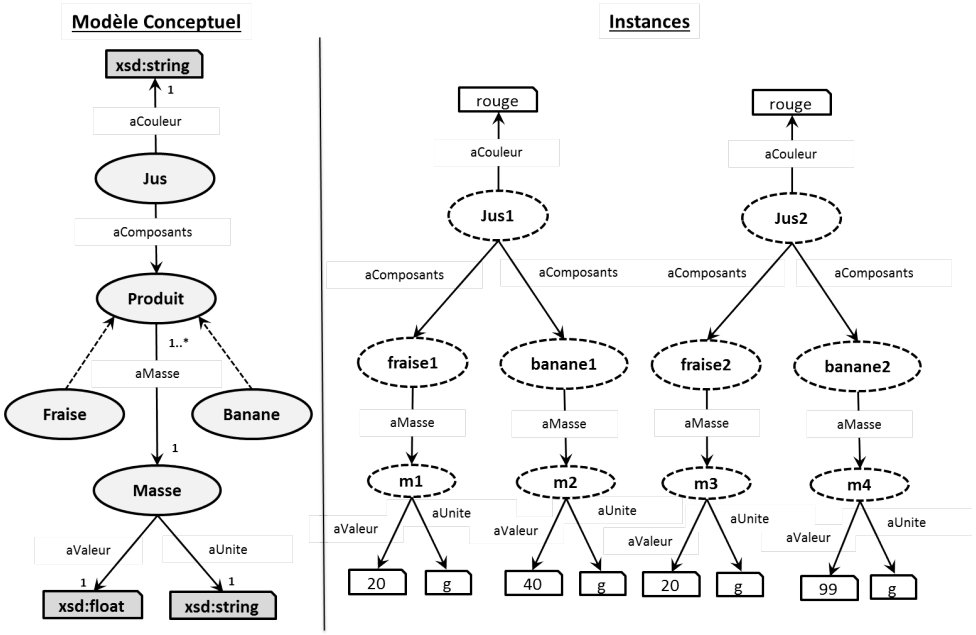


FIGURE 1 – Un extrait d'ontologie O, deux instances jus1 et jus2 de la classe cible *Jus*.

instances de jus de fruit $jus1$ et $jus2$ sont identiques est celui où l'on considère tous les produits

composant le jus, et pour chaque produit une masse dont la description est réduite à l'unité de mesure, mais sans considérer leur quantité. Dans un deuxième contexte, la masse de Fraise est considérée (avec sa valeur et son unité) et l'on observera la présence de banane mais on ne considérera pas la masse pour la banane.
Nous nous intéressons à la recherche de contextes communs correspondant à un sous-ensemble de classes et où pour chaque classe un sous-ensemble de propriétés de l'ontologie peut être pris en compte [3]. Enfin, certains contextes sont clairement plus pertinents que d'autres (e.g. une valeur de masse sans unité n'a pas de sens). Nous considérons également une recherche de liens d'identités contextuels qui puisse être guidée par certaines connaissances expertes.

3.3 Contextes

L'approche de détection de liens d'identité contextuelle que nous avons développée utilise une notion de contexte global fondée sur la notion de contexte local qui peut être définie pour une classe quelconque de l'ontologie.

Définition (**Contexte Local**). Étant donnée une classe c, un contexte local peut être défini par $\pi(c) = (c, DP^c, OP^c)$ où DP^c est un ensemble de propriétés de type *owl*:*DataTypeProperty* telles que c apparaît en domaine, ensemble auquel peuvent s'ajouter la propriété rdf:Type ainsi que des propriétés d'annotations (e.g., *rdfs*:*label*). OP^c est un ensemble de propriétés de type *owl*:*ObjectProperty* telles que c apparaît en domaine ou en co-domaine (notée alors op^{-1}).

Par exemple, un des contextes locaux de la classe $Masse$ présenté en Figure 1 est :
$\pi(Masse) = (Masse, \{rdf\ :type, aValeur, aUnite\}, \{aMasse^{-1}\})$

Définition (**Relation d'ordre sur les contextes locaux**). Soient $\pi_1(c)$ et $\pi_2(c)$ deux contextes locaux pour la classe c. Le contexte $\pi_1(c)$ est dit plus spécifique que $\pi_2(c)$, noté $\pi_1(c) \leqslant \pi_2(c)$, si et seulement si : $DP_2^c \subseteq DP_1^c$ et $OP_2^c \subseteq OP_1^c$.

Par exemple, soient deux contextes locaux de la classe Masse π_1 et π_2 tels que :
$\pi_1(Masse) = (Masse, \{rdf\ :type, aValeur, aUnite\}, \{aMasse^{-1}\})$ et
$\pi_2(Masse) = (Masse, \{rdf\ :type, aValeur\}, \emptyset)$, on a alors $\pi_1(Masse) \leqslant \pi_2(Masse)$.
L'ensemble des contextes locaux d'une classe c peut être représenté par un treillis de contextes locaux noté $T(c)$ qui est composé de 2^n contextes locaux, où n est le nombre de propriétés.

Pour réduire le nombre de contextes locaux à considérer (et donc le nombre de contextes globaux), nous prenons en compte des connaissances expertes, quand elles sont disponibles, concernant l'inutilité ou la nécessité de certaines propriétés ainsi que des informations sur l'importance de la cooccurrence de certaines d'entre elles. Plus précisément, pour chaque classe, un expert peut alimenter trois types de listes :

– $unwantedProps$: la liste des propriétés non pertinentes, qui n'interviendront pas dans

3. Il ne s'agit pas ici de calculer le graphe le plus spécifique partagé par deux instances de Jus dans lequel les descriptions de classes pourraient varier selon les instances considérées (e.g. les propriétés prises en compte pour comparer les instances de la classe Masse ne doivent pas varier selon que l'on compare la masse de la banane ou celle de la fraise)

le calcul des liens d'identité, soit parce qu'il s'agit d'une propriété non structurée (textuelle), ou parce que ses valeurs sont très hétérogènes ou encore parce que ses variations ne sont pas significatives compte tenu de l'utilisation des liens d'identité par l'expert (e.g. la couleur du jus de fruit). Ainsi, si l'on a ajouté la propriété $rdf:type$ à $unwantedProps$ avant la construction de $T(Masse)$, le contexte le plus spécifique sera $\pi_k(Masse) = (Masse, \{aValeur, aUnite\}, \{aMasse^{-1}\})$ et $T(Masse)$ contiendra seulement 8 contextes locaux.

– $coProps$: ensembles de propriétés devant être prises en compte ensemble, si elles sont considérées. Par exemple, une valeur de masse n'ayant pas de sens sans son unité de mesure, le couple $\{aValeur, aUnit\}$ peut être ajouté par l'expert à $coProps$. Dans ce cas, $T(Masse)$ se réduira à quatre contextes locaux possibles.

– $necProps$: ensemble des propriétés essentielles pour la comparaison des instances quel que soit le contexte considéré. Ainsi, si on ajoute $\{rdfType, aValeur, aUnit\}$ à $necProps$, $\pi_k(Masse) = (Masse, \{rdf:type, aValeur, aUnite\}, \{\emptyset\})$ sera le contexte le moins spécifique de $T(Masse)$.

Définition (**Contexte Global**). Etant donnée une classe cible $cbl \in \mathcal{C}$, un contexte global $\Pi(cbl)$ est un sous-graphe connexe G de $\mathcal{G}$ contenant cbl, tel que : $\Pi(cbl) = \bigcup_{c_k \in C_G} \pi(c_k)$

Dans l'exemple présenté en figure 1, il existe différents contextes globaux dans lesquels les deux individus de la classe cible ***Jus*** sont identiques. On en citera deux :

– $\Pi_a(Jus) = \{(Jus, \{rdf:type\}, \{aComposants\}),$
$(Fraise, \{rdf:type\}, \{aComposants^{-1}\}), (Banane, \{rdf:type\}, \{aComposants^{-1}\})\}$

– $\Pi_b(Jus) = \{(Jus, \{rdf:type, aCouleur\}, \{aComposants\}),$
$(Fraise, \{rdf:type\}, \{aComposants^{-1}, aMasse\},$
$(Banane, \{rdf:type\}, \{aComposants^{-1}\}, (Masse, \{rdf:type, aValeur, aUnite\}, \emptyset)\}$

Les contextes globaux que nous considérons ne contiendront pas deux classes c_1 et c_2, telles que $c_1 \subseteq c_2$. Cette contrainte permet d'éviter de construire des contextes globaux ne respectant pas les mécanismes d'héritage des propriétés. Pour cela nous sélectionnons les classes les plus générales pour lesquelles des instances ont été typées directement. Par exemple, s'il existe des instances directes de la classe $Produit$, les contextes globaux ne contiendront plus de contextes locaux définis pour les classes $Fraise$ et $Banane$ qui sont plus spécifiques. Les instances de ces classes seront représentées par le contexte local de la classe $Produit$. Le contexte global est alors plus abstrait et plus contraint puisqu'on ne pourra pas distinguer les contextes locaux de la classe $Fraise$ et de la classe $Banane$, comme pour le contexte global Π_b.

Définition (**Relation d'ordre sur les contextes globaux**). Soient $\Pi_1(cbl)$ et $\Pi_2(cbl)$ deux contextes globaux pour la classe cbl. Le contexte global $\Pi_1(cbl)$ est dit plus spécifique que $\Pi_2(cbl)$, noté $\Pi_1(cbl) \leqslant \Pi_2(cbl)$, si et seulement si : $\forall \pi_i(c) \in \Pi_2(cbl), \exists \pi_j(c) \in \Pi_1(cbl)$ tel que $\pi_j(c) \leqslant \pi_i(c)$.

Dans l'exemple ci-dessus, $\Pi_b(jus) \leqslant \Pi_a(jus)$, puisque $\pi_b(Jus) \leqslant \pi_a(Jus)$, $\pi_b(Fraise) \leqslant \pi_a(Fraise)$ et $\pi_b(Banane) \leqslant \pi_a(Banane)$.

3.4 Relation d'identité contextuelle

Une relation d'identité contextuelle est valide dans un contexte global si toutes les valeurs littérales (à une équivalence près) et toutes les instances appartenant à ce contexte sont identiques. Cette relation peut être exprimée de la façon suivante :

Définition (**Relation d'identité contextuelle**). Soient (i_1, i_2) deux instances d'une classe cible cbl. Soit $\Pi(cbl) = \{\pi_1(c_1), \ldots, \pi_n(c_n)\}$ un contexte global pour la classe cbl. Deux instances (i_1, i_2) sont liées par une relation d'identité dans le contexte global $\Pi(cbl)$, notée $identiConTo_{<\Pi(cbl)>}(i_1, i_2)$, si et seulement si les sous-graphes RDF décrivant i_1 et i_2 et obtenus en utilisant la partie de l'ontologie représentée dans le contexte global, sont identiques, au renommage d'URI près et à une réécriture de valeurs littérales près.

Les relations d'identités ne seront représentées que pour les contextes les plus spécifiques et pourront être générées pour des contextes plus généraux si besoin. Soient $\Pi_1(cbl)$ et $\Pi_2(cbl)$ deux contextes globaux d'une classe cible cbl. Si $\Pi_1(cbl) \leqslant \Pi_2(cbl)$ alors $identiConTo_{<\Pi_1(cbl)>}(i_1, i_2) \Rightarrow identiConTo_{<\Pi_2(cbl)>}(i_1, i_2)$

4 DECIDE - Un algorithme de détection de liens d'identité contextuels

Le but de l'algorithme $DECIDE$ (DEtection of Contextual IDEntity) est de déterminer pour chaque couple d'instances (i_1,i_2) $\in I \times I$ d'une classe cible cbl fixée par l'utilisateur, l'ensemble des contextes globaux les plus spécifiques pour lesquels la relation $identiConTo$ est vraie. $DECIDE$ prend en paramètres la base de connaissances $\mathcal{B}$, la classe cible cbl, et les trois listes de contraintes des experts $unwantedProps$, $coProps$ et $necProps$ si elles existent. La construction des contextes globaux se déroulent en trois étapes :

- Construction de la liste $Cdep$ des classes les plus générales du graphe connexe maximal de cbl qui comportent des instances directement typées par ces classes.
- Pour chaque classe $c \in Cdep$, construction des treillis de contextes locaux $T(c)$ pertinents en tenant compte des listes $unwantedProps$, $coProps$ et $necProps$ définies par les experts.
- Pour chaque couple d'instances (i_1,i_2) de cbl, appel de la fonction $IdentiConMax$ détaillée dans l'algorithme 1 qui calcule l'ensemble des contextes globaux les plus spécifiques ($CGSet$).

Pour chaque couple d'instances (i_1,i_2), $IdentiConMax$ retourne l'ensemble des contextes globaux les plus spécifiques, dans lesquels les instances du couple (i_1,i_2) sont identiques. Cette fonction effectue un parcours en profondeur d'abord des propriétés décrivant les instances à comparer et construit au fur et à mesure les contextes globaux les plus spécifiques.

Algorithme 1 : identiConToMax

Inputs :
- cbl : la classe cible
- $Cdep$: l'ensemble des classes du graphe connexe maximal de cbl
- (i_1, i_2) : une paire d'instances de la classe c

Output : $CGSet$: l'ensemble des contextes globaux les plus spécifiques concernant le couple (i_1, i_2)

$CGSet \quad \emptyset$; $\Pi(cbl) \quad \emptyset$; $fileCP.add(\Pi(cbl))$

while $fileCP$ *is not empty* **do**

  $\Pi(cbl)^{cour} \quad fileCP.getNextElement()$

  $dejaVu \leftarrow \emptyset$; $propSrc \leftarrow nil$; $cSrc \leftarrow nil$

  $\Pi(cbl)^{cour} = compareCouple(i_1, i_2, dejaVu, propSrc, cSrc, \Pi(cbl)^{cour})$

  $CGSet \quad CGSet \cup \Pi(cbl)^{cour}$

Plus précisément, l'algorithme commence avec un contexte global courant vide $\Pi(cbl)^{cour}$ qui sera enrichi par les contextes locaux au fur et à mesure de l'exploration. Pour ce contexte, on appelle la fonction $compareCouple$ qui permet de comparer un couple d'instances d'une même classe c. Une file de contextes globaux partiels, notée $fileCP$, contient les contextes globaux alternatifs qui seront détectés lors de l'exploration et qui ne sont pas encore traités.

La fonction $compareCouple$, décrite dans l'algorithme 2, permet d'enrichir un contexte courant pour trouver un contexte plus spécifique dans lequel les instances du couple (i_1,i_2) sont identiques. Les données étant représentées sous la forme d'un graphe pouvant comporter des cycles, une liste de paires d'instances déjà explorées nommée $dejaVu$ est maintenue afin d'éviter de recalculer les contextes identiques pour ces paires. La fonction $comparerCouple$ prend en paramètre la liste $dejaVu$, la propriété objet source $propSrc$ qui a permis d'atteindre le couple (i_1,i_2) à comparer, la classe source $cSrc$ et le contexte global en cours $\Pi(cbl)^{cour}$ ($dejaVu$, $propSrc$ et $cSrc$ seront réinitialisés à nil pour chaque nouveau contexte courant exploré). Cette fonction détermine les propriétés de type *DataTypeProperty* et les propriétés d'annotation dont les valeurs sont identiques (noté $DPeg$)[4]. Elle détermine également l'ensemble des propriétés objets $OPcom$ instanciées pour i_1, et i_2. La fonction $possibleContext$ va ensuite rechercher dans le treillis $T(c)$ si ce contexte local $\pi(c)^{cour}$ est possible (compte tenu des connaissances expert). S'il n'existe pas un tel contexte, le contexte local de $T(c)$ le plus spécifique généralisant $\pi(c)^{cour}$ est retourné. Si ce contexte retourné est le contexte vide, la fonction s'arrête et retourne le contexte global en cours $\Pi(cbl)^{cour}$. Si le contexte retourné de la fonction $possibleContext$ n'est pas vide, 4 cas sont possibles :

1 – Il n'existe pas de contexte local déjà défini pour c dans $\Pi(cbl)^{cour}$, on ajoute alors $\pi(c)^{cour}$ à $\Pi(cbl)^{cour}$. Ensuite, la fonction $compareObjProperties$, est appelée pour comparer tous les ressources reliées au couple (i_1, i_2). Par exemple, si i_1=$Jus1$ et i_2=$Jus2$, la fonction $compareObjProperties$ va comparer les couples $(fraise1, fraise2)$ puis $(banane1, banane2)$.

2 – Il existe un contexte local identique $\pi(c)^{existant}$ déjà défini pour c. On appelle alors directement $compareObjProperties$.

3 – Il existe un contexte local pour c dans $\Pi(cbl)^{cour}$ qui est plus général que $\pi(c)^{cour}$. Dans

4. Une mesure de similarité peut être utilisée et différer suivant chaque propriété

ce cas, on crée un nouveau contexte global $\Pi(cbl)^{nouv}$ contenant le contexte local $\pi(c)^{cour}$, et on l'ajoute à la file $fileCP$ qu'il restera à explorer. Ensuite, comme $\pi(c)^{cour} \leqslant \pi(c)^{existant}$, on doit vérifier si les instances ont bien les mêmes valeurs pour les *objectProperties* de $\pi(c)^{existant}$ en appelant la fonction $compareObjProperties$.

4 – Dans tous les autres cas, on crée un nouveau contexte global $\Pi(cbl)^{nouv}$ contenant le contexte local $\pi(c)^{cour}$, et on l'ajoute à la file $fileCP$ pour qu'il soit exploré dans une prochaine itération de l'algorithme 1. Dans le contexte courant, comme on est sûr que le couple n'est pas identique en considérant $\pi(c)^{existant}$, le contexte de la classe source src $\pi(src)^{existant}$ est remplacé par un nouveau contexte local $\pi(src)^{nouv}$ ne contenant pas la propriété $propSrc$ dans $\Pi(cbl)^{cour}$.

Algorithme 2 : compareCouple

Inputs :
- (i_1, i_2) : une paire d'instances
- $dejaVu$: la liste des couples d'instances déjà traités
- $propSrc$: l'object property source
- $cSrc$: la classe source (i.e., le domaine ou le co-domaine de $propSrc$)
- $\Pi(cbl)^{cour}$: le contexte global courant de la classe cible.

Output : $\Pi(cbl)^{cour}$: le contexte global courant de la classe cible, mis à jour.

$dejaVu \leftarrow dejaVu \cup (i_1, i_2)$;
$c \leftarrow$ la classe commune de i_1 et i_2 tel que $c \in Cdep$;
$OPcom \leftarrow$ les propriétés objets communes de i_1 et i_2;
$DPeg \leftarrow$ les data properties, rdftype et propriétés d'annotation ayant mêmes valeurs pour i_1 et i_2;
$\pi(c)^{cour} \leftarrow possibleContext(c, DPeg, OPcom)$
if $\pi(c)^{cour}$ *!=* $\emptyset$ **then**
  if *(Il n'existe pas de contexte local de c dans* $\Pi(cbl)^{cour}$*)* **then**
    $\Pi(cbl)^{cour}.add(\pi(c)^{cour})$
    $compareObjProperties(i_1, i_2, c, dejaVu, propSrc, cSrc, \Pi(cbl)^{cour})$;
  else
    $\pi(c)^{existant} \leftarrow getLocalContext(c, \Pi(cbl)^{cour})$
    if $\pi(c)^{existant} == \pi(c)^{cour}$ **then**
      $compareObjProperties(i_1, i_2, c, dejaVu, propSrc, cSrc, \Pi(cbl)^{cour})$;
    else
      $\Pi(cbl)^{nouv} \leftarrow \emptyset$
      $\Pi(cbl)^{nouv}.add(\pi(c)^{cour})$
      if $\Pi(cbl)^{nouv}$ *n'existe pas dans* $fileCP$ **then**
        $fileCP \leftarrow fileCP.add(\Pi(cbl)^{nouv})$
      if $\pi(c)^{cour} \leqslant \pi(c)^{existant}$ **then**
        $compareObjProperties(i_1, i_2, c, dejaVu, propSrc, cSrc, \Pi(cbl)^{cour})$;
      else
        $\pi(cSrc) \leftarrow getLocalContext(cSrc, \Pi(cbl)^{cour})$
        $\pi(cSrc)^{nouv} \leftarrow possibleContext(cSrc, \pi(cSrc).DPeg, \pi(cSrc).OPcom - propSrc)$
        $\Pi(cbl)^{cour} \leftarrow replaceLocalContext(cSrc, \Pi(cbl)^{cour}, \pi(cSrc)^{nouv})$
return $\Pi(cbl)^{cour}$;

La fonction $compareObjProperties$ appelée dans $identiConToMax$ permet de comparer l'ensemble des instances reliées au couple (i_1, i_2) pour chacune des propriétés p du contexte local. Les listes des instances liées à i_1 et i_2 sont examinées classe par classe et les propriétés sont supposée être localement complètes pour chacune des classes (e.g. si l'on connait certains fruits composants un jus, nous supposons qu'ils sont tous représentés). Pour chacune des classes, afin

	Echantillon	***D. Réelles***
#Faits	24 458	1 269 624
#Classes	4 635	4 743
#Instances	876	272 724
#ObjectProperties	50	50
#DataProperties	11	11
#Annot.Properties	25	25

	Echantillon	***D. Réelles***
# Instances (cible)	22	220
# Paires d'instances	231	24 090
# Classes (graphe)	128	306
# Instances (graphe)	606	90 384
# Contextes Globaux	8	30
# Liens d'identité	252	25 189
Temps d'exécution	9 s	372 s

TABLE 1 – Taille des Datasets et Résultats

de limiter le nombre de paires à examiner en cas de multivaluation, on vérifie tout d'abord que le nombre de valeurs de p est identique, puis on utilise une heuristique, qui permet de ne former que les couples (i'_1, i'_2) qui partagent le plus de valeurs de *data properties*. Après avoir formé la liste des meilleures paires, la fonction rappelle la fonction $compareCouple$ pour chacune de ces paires afin de trouver les contextes les plus spécifiques dans lesquels ces paires sont identiques.

L'algorithme $DECIDE$ appliqué sur l'exemple de la figure 1 permet de trouver les contextes globaux les plus spécifiques suivants pour $Jus1$ et $Jus2$:
$\Pi_a(Jus) = \{(Jus, \{rdf : type, aCouleur\}, \{aComposants\}), (Fraise, \{rdf : type\}, \{aComposants^{-1}, aMasse\}), (Banane, \{rdf : type\}, \{aComposants^{-1}\}), (Masse, \{rdf : type, aValeur, aUnite\}, \{aMasse^{-1}\})\}$ et $\Pi_b(Jus) = \{(Jus, \{rdf : type, aCouleur\}, \{aComposants\}), (Fraise, \{rdf : type\}, \{aComposants^{-1}, aMasse\}), (Banane, \{rdf : type\}, \{aComposants^{-1}, aMasse\}), (Masse, \{rdf : type, aUnite\}, \{aMasse^{-1}\})\}$

5 Expérimentations

Notre approche a été évaluée sur des données scientifiques relatives au domaine de l'agro-alimentaire. Nous avons exploité la version 1.4 de l'ontologie PO^2 (Ibanescu *et al.*, 2016) qui décrit des processus de transformations [5]. Dans le cadre du projet INRA CellExtraDry, PO^2 a été enrichie par une partie d'Agrovoc et la version de PO^2 a été peuplée par 10 processus de transformation de micro-organismes (levures) suivant 20 itinéraires où chaque itinéraire représente des séquences d'étapes de transformation (séchage, chauffage ...). Les données réelles ne pouvant être mises à disposition, pour des raisons de confidentialité, un échantillon de ces données, décrivant un seul processus de stabilisation et dont certaines valeurs ont été modifiées, est accessible avec le code en $Java$ de l'algorithme et ses résultats sur `https://github.com/raadjoe/DECIDE`. La taille des données réelles du projet et celle de l'échantillon sont décrites dans la Table 1. Cette première expérimentation a eu pour but d'observer le nombre de contextes distincts et le nombre de liens d'identité contextuels pouvant être générés dans une ontologie décrivant des données scientifiques. Dans cette expérimentation, une égalité stricte des valeurs littérales a été utilisée (à une normalisation près).

L'algorithme $DECIDE$ a été appliqué pour chacun de ces datasets afin de découvrir les contextes les plus spécifiques pour lesquels les individus de la classe cible $Mixture$ sont identiques. Une instance de cette classe représente un ensemble de produits sur lequel est appliqué une étape d'un processus de transformation. La taille de la classe cible pour chaque dataset, la

5. L'ontologie core de PO^2 est accessible sur `http://agroportal.lirmm.fr/ontologies/PO2`

taille du graphe connexe maximal (i.e. classes et individus pouvant être atteints depuis la classe cible), et les résultats de l'algorithme sont présentés dans la Table 1. L'algorithme a été exécuté sur une machine de 8GB de RAM, ayant un processeur Intel Core 4× 2.6GHz (Windows 10).

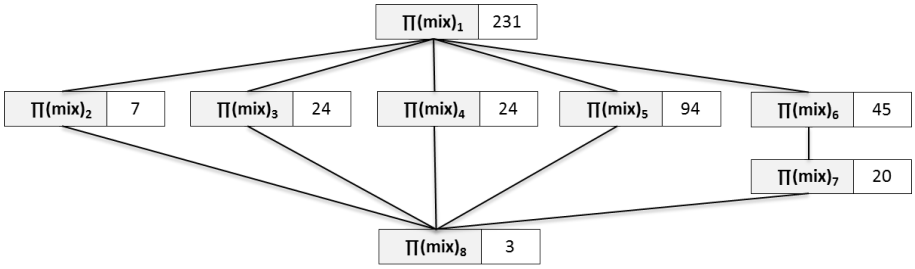


FIGURE 2 – La hiérarchie des 8 contextes globaux les plus spécifiques de la classe Mixture.

Les 220 instances de la classe cible $Mixture$ des données réelles, forment 24 090 couples d'individus candidats. 30 contextes globaux générés par $DECIDE$ suffisent pour représenter les contextes les plus spécifiques dans lesquels ces couples sont identiques (8 pour l'échantillon). Le nombre de liens d'identité (25 189), plus élevé que le nombre de couples d'individus, montre que certains couples sont identiques dans plusieurs contextes globaux non comparables. La hiérarchie des 8 contextes globaux créés dans l'échantillon est représentée dans la Figure 2. Le nombre qui apparait à droite de chaque contexte global dans cette figure représente le nombre de couples qui sont identiques dans ce contexte (après inférence). Le contexte $\Pi(mix)_1 = \{(Mixture, \{rdf : type\})\}$, qui est le contexte global le moins spécifique de tous les contextes générés, est un contexte dans lequel tous les couples sont identiques (231 pour l'échantillon). Le contexte le plus spécifique qui a été généré ($\Pi(mix)_8$), contient 11 contextes locaux, et peut garantir une identité plus forte. Un couple de mixtures identiques dans ce contexte signifie qu'elles sont de même nature (e.g humides) et partagent les mêmes quantités et unités de mesure de leurs composants : eau, silicone, levure, glucose, et cystéine.

En ré-appliquant l'algorithme $DECIDE$, et en prenant compte cette fois d'une seule connaissance experte qui indique qu'une valeur de masse n'a pas de sens sans son unité de mesure et vice-versa, les contextes locaux de 17 contextes globaux parmi 30 des données réelles, et 5 parmi 8 contextes globaux de l'échantillon étaient affectés. Ce qui indique que l'ajout d'une seule connaissance au début de l'algorithme a permis d'améliorer la sémantiques d'environ 60% des contextes globaux générés, sans affectant le nombre des contextes pour les deux datasets.

6 Conclusion

Nous proposons une approche de détection de liens d'identité contextuels permettant de générer pour chaque couple instance d'une classe cible, l'ensemble des contextes les plus spécifiques dans lesquels ces instances sont identiques. Une première expérimentation a été réalisée pour des données réelles décrivant des expérimentations scientifiques sur la levure, menées à l'INRA. Cette expérimentation a montré que différents contextes à différents niveaux d'abstraction pouvaient être découverts en un temps raisonnable.

Nous souhaitons maintenant montrer les résultats obtenus aux experts de domaine pour connaître l'intérêt des différents contextes et recueillir le maximum de contraintes permettant

d'élaguer l'espace de recherche. Cette collecte pourrait être facilité par la création d'un outil permettant d'interagir avec les experts du domaine. Nous souhaitons ensuite utiliser ces liens pour faire prédire des valeurs manquantes à différents niveaux de confiance selon le niveau de spécificité du contexte d'identité. Il nous envisageons également de découvrir relations de causalité entre certains faits décrivant les expériences et ceux décrivant les résultats des observations en considérant l'identité contextuelle des autres facteurs expérimentaux.

Remerciements

Ce travail est soutenu par le Center for Data Science, financé par IDEX Paris-Saclay, ANR-11-IDEX-0003-02.

Références

BEEK W., SCHLOBACH S. & HARMELEN F. (2016). A contextualised semantics for owl : Sameas. In *Proceedings of the 13th International Conference on The Semantic Web. Latest Advances and New Domains - Volume 9678*, p. 405–419, New York, NY, USA : Springer-Verlag New York, Inc.

DE MELO G. (2013). Not quite the same : Identity constraints for the web of linked data. In M. DESJARDINS & M. L. LITTMAN, Eds., *AAAI* : AAAI Press.

DING L., FININ T., SHINAVIER J. & MCGUINNESS D. L. (2010). owl :sameAs and linked data : An empirical study. In *In The Semantic Web - ISWC*, p. 145–160.

FERRARA A., NIKOLOV A. & SCHARFFE F. (2011). Data linking for the semantic web. *Int. J. Semantic Web Inf. Syst.*, **7**(3), 46–76.

HALPIN H., HAYES P. J., MCCUSKER J. P., MCGUINNESS D. L. & THOMPSON H. S. (2010). When owl :sameas isn't the same : An analysis of identity in linked data. In P. F. PATEL-SCHNEIDER, Y. PAN, P. HITZLER, P. MIKA, L. ZHANG, J. Z. PAN, I. HORROCKS & B. GLIMM, Eds., *The Semantic Web – ISWC 2010 : 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I*, p. 305–320, Berlin, Heidelberg : Springer Berlin Heidelberg.

HALPIN H., HAYES P. J. & THOMPSON H. S. (2015). When owl : sameas isn't the same redux : towards a theory of identity, context, and inference on the semantic web. In *International and Interdisciplinary Conference on Modeling and Using Context*, p. 47–60 : Springer.

IBANESCU L., DIBIE J., DERVAUX S., GUICHARD E. & RAAD J. (2016). Poˆ 2-a process and observation ontology in food science. application to dairy gels. In *Metadata and Semantics Research : 10th International Conference, MTSR 2016, Göttingen, Germany, November 22-25, 2016, Proceedings*, p. 155–165 : Springer.

JAFFRI A., GLASER H. & MILLARD I. (2008). Uri disambiguation in the context of linked data. In C. BIZER, T. HEATH, K. IDEHEN & T. BERNERS-LEE, Eds., *Linked Data on the Web - LDOW*, volume 369 of *CEUR Workshop Proceedings* : CEUR-WS.org.

MILES A. & BECHHOFER S. (2009). Skos simple knowledge organization system reference. w3c recommendation 18 august 2009.

PAPALEO L., PERNELLE N., SAÏS F. & DUMONT C. (2014). Logical detection of invalid sameas statements in RDF data. In *Knowledge Engineering and Knowledge Management - 19th International Conference, EKAW 2014, Linköping, Sweden, November 24-28, 2014. Proceedings*, p. 373–384.

PATEL-SCHNEIDER P. F., HAYES P. & HORROCKS I. (2004). *OWL Web Ontology Language Semantics and Abstract Syntax Section 5. RDF-Compatible Model-Theoretic Semantics*. Rapport interne, W3C.

Un jeu de données d'évaluation de correspondances complexes entre ontologies

Elodie Thiéblin, Ollivier Haemmerlé, Nathalie Hernandez, Cassia Trojahn

IRIT-UT2J
Institut de recherche informatique de Toulouse, Toulouse, France
nom.prenom@irit.fr

Résumé : Le web de données liées se compose d'entrepôts de données décrits par des ontologies. Ces ontologies hétérogènes ont différents niveaux de disparité (différences terminologiques, de conceptualisation, de modélisation). Les alignements simples font correspondre une entité de l'ontologie source à une entité de l'ontologie cible. Les alignements complexes, complètent les alignements simples en exprimant plus finement les différents types de disparités. Des approches permettant de détecter des correspondances complexes entre ontologies émergent et il n'existe pas encore de jeu de données exhaustif pour les évaluer. Cet article présente un jeu de données d'alignements complexes entre deux paires d'ontologies du domaine de l'organisation de conférences, ainsi qu'une évaluation d'approches d'alignement permettant d'obtenir de telles correspondances.
Mots-clés : Alignements d'ontologies, correspondances complexes, évaluation.

1 Introduction

Le web de données liées (LOD) est composé de nombreux entrepôts de données. Ces données sont décrites par différents vocabulaires (ou ontologies). La diversité des ontologies utilisées sur le LOD est source d'hétérogénéité. Klein (2001) distingue plusieurs niveaux d'hétérogénéité entre ontologies : les différences de conceptualisation (sur le domaine, la portée des ontologies et leur granularité), terminologiques (termes associés aux entités d'une ontologie) et de modélisation (conventions de modélisation et paradigmes utilisés).

L'alignement d'ontologies (Euzenat & Shvaiko (2013)) est une solution à ces problèmes d'hétérogénéité. On distingue deux catégories d'alignements : les alignements simples (composés de correspondances simples) et les alignements complexes (ayant au moins une correspondance complexe). Les correspondances simples sont limitées car elles lient une à une chaque entité d'ontologies. Les correspondances complexes sont le complément des correspondances simples car elles peuvent mieux prendre en compte les différences de modélisation entre ontologies. Ces différences de modélisation peuvent être répertoriées en *patrons de correspondance*. Les approches de détection de correspondances simples (ou approches d'alignement simple) sont relativement nombreuses et de plus en plus performantes (Achichi *et al.* (2016)). Les approches de détection de correspondances complexes sont moins nombreuses même si de nouvelles propositions voient régulièrement le jour (Qin *et al.* (2007); Ritze *et al.* (2009, 2010); Parundekar *et al.* (2012, 2010); Walshe (2014)). Parmi ces approches, une des plus significatives est basée sur des patrons de correspondance (Scharffe (2009)).

Afin de connaître les atouts et de soulever des pistes d'amélioration d'une approche d'alignement, il est important de l'évaluer. C'est le but de L'OAEI (Ontology Alignment Evaluation Initiative) (Achichi *et al.* (2016)) qui évalue des approches d'alignement sur différents jeux de données. Pour l'instant, ces jeux sont centrés sur les alignements simples.

Dans ce papier, un jeu de données composé de trois ontologies et de correspondances complexes entre deux paires de ces ontologies est proposé. La méthodologie utilisée pour construire ce jeu de données est décrite. Une approche d'alignement basée sur des patrons de correspondance est évaluée sur ce jeu de données. Nous présentons le résultat de l'évaluation. Cette évaluation donne lieu à une discussion sur les types de correspondances complexes et les pistes d'amélioration du jeu de données proposé.

2 Prérequis et travaux liés

2.1 Définitions

Un alignement est un ensemble de correspondances entre une ontologie source $o1$ et une ontologie cible $o2$. On différencie deux types de correspondances :

— Les *correspondances simples* mettent en relation deux entités atomiques de deux ontologies (ici exprimées en logique du premier ordre). Par exemple $\forall x$, *o1#Paper*(x) $\equiv$*o2#Article*(x) est une correspondance simple, *Paper* et *Article* étant deux classes de $o1$ et $o2$, respectivement. On parle de correspondances de cardinalité 1:1 pour désigner les correspondances simples.

— Les *correspondances complexes* permettent d'exprimer des formules logiques entre entités de $o1$ et $o2$. Par exemple $\forall x$, *o1#Accepted_Paper*(x) $\equiv$ $\exists y$,*o2#hasDecision*(x, y) $\wedge$ *o2#Acceptance*(y) est une correspondance complexe car au moins un des membres de l'équivalence est une construction logique d'entités. Les correspondances complexes ont pour cardinalité 1:n, n:1 ou n:m suivant le nombre d'entités et de constructeurs présents de chaque côté de la relation.

La relation d'une correspondance peut être une équivalence ($\equiv$), une relation de subsomption ($\geq$, $\leq$). Un alignement est dit complexe quand au moins une de ses correspondances est complexes. Dans la suite de cet article, les ontologies sont représentées comme suit : les rectangles sont des classes ; une flèche marquée de ($\geq$) entre deux classes marque une relation de subsomption ; un arc marqué ($\perp$) entre deux classes représente leur disjonction ; une flèche verte en trait continu entre deux classes est une propriété sur les objets et les classes qu'elle lie son domaine et co-domaine ; une flèche verte en trait pointillés est une propriété sur les données, elle lie son domaine à son type de données. Entre les entités de deux ontologies, les correspondances simples sont représentées en rouge et la relation de la correspondance est indiquée sur le lien. Les correspondances complexes ne sont pas représentées graphiquement. La figure 1 représente les fragments d'ontologies utilisés dans les correspondances ci-dessus.

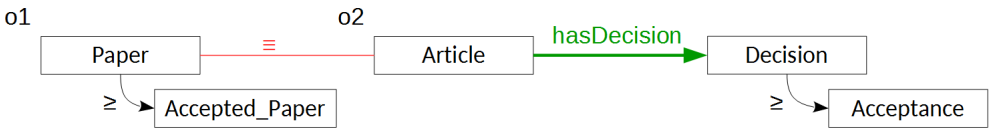


FIGURE 1 – Deux fragments d'ontologies

2.2 Patrons de correspondance

Scharffe (2009) définit les patrons de correspondance comme des solutions pour identifier des types de disparités récurrentes au niveau de la modélisation de deux ontologies. Il propose une librairie de 36 patrons de correspondances qui peuvent être assemblés en patrons composés. Par soucis de place, nous ne présentons que deux patrons représentatifs de ceux utilisés par les

approches. Le patron *Class by Attribute Type* (CAT) décrit les correspondances de la forme $\forall x, A(x) \equiv \exists y, b(x,y) \wedge C(y)$. Autrement dit, la classe A est représentée par une restriction de type C sur l'attribut (propriété sur les objets) b. Par exemple, $\forall x$, *Accepted_Paper(*x*)* $\equiv \exists y$, *hasDecision(*x, y*)* $\wedge$ *Acceptance(*y*)* est une correspondance CAT. Le patron *Class by Attribute Value* (CAV) définit qu'une classe A peut être équivalente à un attribut b dont l'objet est restreint à une valeur donnée v. Une correspondance CAV a pour forme $\forall x, A(x) \equiv b(x,v)$. Par exemple, $\forall x$, *Accepted_Paper(*x*)* $\equiv$ *is_accepted(*x*,true)* est une correspondance CAV.

2.3 Approches d'alignement complexe entre ontologies

Les approches d'alignement complexe entre ontologies émergent. Bien qu'il existe plusieurs approches d'alignement complexe entre schémas, nous faisons la distinction entre ontologie et schéma (XML, base de données, etc.). Les approches d'alignement de schémas sortent du contexte de l'étude. Certaines approches d'alignement d'ontologies se basent sur des patrons de correspondances (Ritze *et al.* (2009, 2010); Parundekar *et al.* (2010, 2012); Walshe (2014)) tandis que d'autres ne les exploitent pas (Qin *et al.* (2007); Nunes *et al.* (2011)).

Ritze *et al.* (2009, 2010) définissent les conditions nécessaires à la détection de correspondances sur la base de certains patrons. Les conditions définies dans ces approches exploitent un alignement simple fourni par l'utilisateur en entrée du processus. Ces conditions portent notamment sur les labels des entités et la structure (taxonomique et propriétés sur les objets) des ontologies. Ces deux approches diffèrent par les indices linguistiques qu'elles considèrent : Ritze *et al.* (2010) utilise des méthodes linguistiques (pré-traitement, exploitation de relations lexicales,etc.) tandis que Ritze *et al.* (2009) utilise uniquement des similarités entre chaînes de caractères. Aucune des deux approches n'utilise ni ne nécessite des instances.

Les autres approches d'alignement requièrent et utilisent des instances communes à deux bases de connaissance décrites par les ontologies à aligner. Parundekar *et al.* (2010, 2012) proposent des approches fondées sur des patrons et des instances communes. Les correspondances détectées sont de type "attribut-valeur = conjonction d'attribut-valeur" ou "attribut-valeur = attribut-ensemble de valeurs". L'espace de recherche est représenté par un arbre puis "élagué" pour trouver la meilleure correspondance. Walshe (2014) propose une approche également fondée sur les patrons et les instances communes. Son approche se concentre sur la détection de correspondances CAV en utilisant des méthodes de sélection d'attributs. Nunes *et al.* (2011) n'utilise pas de patron de correspondance mais utilise les instances pour chercher des combinaisons de propriétés utilisant des fonctions de transformation (concaténation de chaînes de caractères, etc.) par programmation génétique. Qin *et al.* (2007) ne se fonde pas sur les patrons de correspondance. Son approche explore les instances communes à deux bases de connaissances pour en extraire des ensembles de prédicats qui ont un nombre d'occurrences supérieur à un certain seuil.

Une des caractéristiques récurrente dans les approches d'alignement est l'utilisation de patrons de correspondance, ce qui sera exploité dans la suite de ce papier.

2.4 Evaluation des approches d'alignement complexe

Il existe de nombreux jeux de données pour évaluer les approches d'alignement simple. Ces jeux de données ont chacun une particularité suivant le type d'ontologies à aligner : ontolo-

gies volumineuses, sur un domaine particulier (e.g. médical), ontologies légères ou lourdes, etc. L'OAEI (Achichi *et al.*, 2016) utilise certains de ces jeux de données pour évaluer les approches d'alignement simple. Les approches d'alignement complexe citées précédemment ont été évaluées à la main essentiellement sur les sorties qu'elles produisent.

Les approches de Ritze *et al.* (2009, 2010) prennent en entrée les ontologies conférence de l'OAEI et les correspondances complexes générées ont été évaluées en terme de précision. Ritze *et al.* (2009) a aussi été évaluée à partir des ontologies Benchmark de l'OAEI.

Les approches d'alignement proposées par Parundekar *et al.* (2010, 2012) ont trouvé de nombreuses correspondances entre divers entrepôts de données du LOD [1]. Etant donné le grand nombre de correspondances détectées, un sous-ensemble a été évalué à la main. La précision et le rappel ont été calculés sur un sous-ensemble de *GeoNames* et *DBpedia* portant sur les pays. Parmi toutes les correspondances trouvées et correctes, on peut se demander si toutes sont nécessaires. En effet, certaines peuvent notamment être décomposées en correspondances simples. Par exemple (avec *dbo* le préfixe pour l'ontologie de *DBpedia* et *lgdo* celui de l'ontologie de *LinkedGeoData*), $\forall$ *x,y, rdf#type(x,dbo#Populated_Place)* $\equiv$ *rdf#type(x,lgdo#Place)* pourrait être remplacée par la correspondance simple $\forall$ *x, dbo#Populated_Place(x)* $\equiv$ *lgdo#Place(x)*.

Walshe (2014) crée des ontologies synthétiques (ensembles de classes) à partir d'instances de DBpedia partageant une paire "attribut-valeur" et d'autres classes ne pouvant pas apparaître dans une correspondance CAV avec DBpedia. La précision est évaluée sur les correspondances produites suivant le nombre d'instances communes.

Qin *et al.* (2007) propose les résultats de son approche [2] évalués à la main entre deux bases de connaissances. Parmi les 9 règles proposées en référence, seules deux ne sont pas décomposables en correspondances simples.

Les jeux de données d'alignements complexes existants ne sont pas réutilisables pour toutes les approches d'alignement complexe car ils ont été créés pour (ou par) une approche donnée. À notre connaissance, il n'existe aucun jeu de données d'alignements complexes entre ontologies complet et réutilisable pour évaluer les approches d'alignement complexe.

3 Méthodologie de création de jeu de données d'alignement complexe

Nous décrivons la méthodologie employée pour constituer le jeu de données proposé. Cette méthodologie étant appliquée à la main, elle n'est pas adaptée à des ontologies de grande taille.

Le but de cette méthodologie est de traduire chaque entité de l'ontologie source $o1$ en utilisant les entités de l'ontologie cible $o2$ et d'assurer que toutes ces correspondances sont cohérentes. La méthodologie se focalise donc sur la découverte de correspondances d'équivalence 1:n permettant une certaine exhaustivité (par rapport aux découvertes de correspondances n:m). L'alignement créé se veut exhaustif pour les correspondances 1:n d'équivalence.

1. `http://www.isi.edu/integration/data/UnionAlignments/`
   `http://www.isi.edu/integration/data/LinkedData/`
2. `http://aimlab-server.cs.uoregon.edu/services/ontomap/`

3.1 Étapes

Avant toute chose, il faut déterminer le but de l'alignement. Le type d'alignement ne sera pas forcément le même qu'il serve à fusionner des ontologies ou qu'il serve de médiateur pour de la réécriture de requêtes. Dans le cas de fusion d'ontologies, l'alignement devra être cohérent ; il ne devra pas contenir de correspondances menant à des inférences inexactes. Pour de la réécriture de requêtes, la cohérence d'un alignement est moins cruciale (Euzenat & Le Duc, 2012). Dans le cadre de cet article, nous cherchons à obtenir un alignement cohérent. La méthodologie s'applique en 5 étapes :

1. Obtenir des correspondances simples entre $o1$ et $o2$. Si cet alignement n'existe pas, les approches de l'état de l'art peuvent être réutilisées.
2. Mettre en correspondance une à une chaque entité de $o1$ (classe, relation et propriété) en se basant sur ses labels, axiomes, définition, contexte, instances (s'il y en a) et l'alignement simple. On cherche si possible une équivalence pour l'entité. Si aucune équivalence n'est trouvée, on cherche une relation de subsomption.
3. Vérifier manuellement les conditions de cohérence pour chaque correspondance produite en fonction de la nature de l'entité source et des patrons impliqués dans la correspondance : classe (section 3.2), relation (propriété sur les objets) (section 3.3) ou propriété (propriété sur les données) (section 3.4).
4. Vérifier la cohérence globale de l'alignement en fusionnant les deux ontologies à l'aide des correspondances.
5. Filtrer les correspondances obtenues pour éviter la redondance (section 3.5).

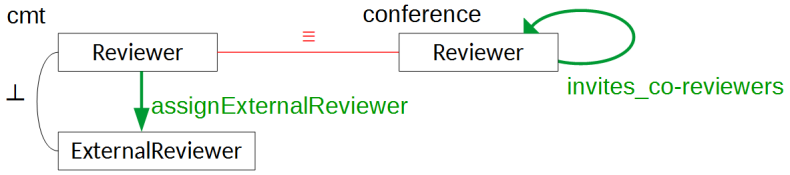


FIGURE 2 – Fragments d'ontologies du dataset

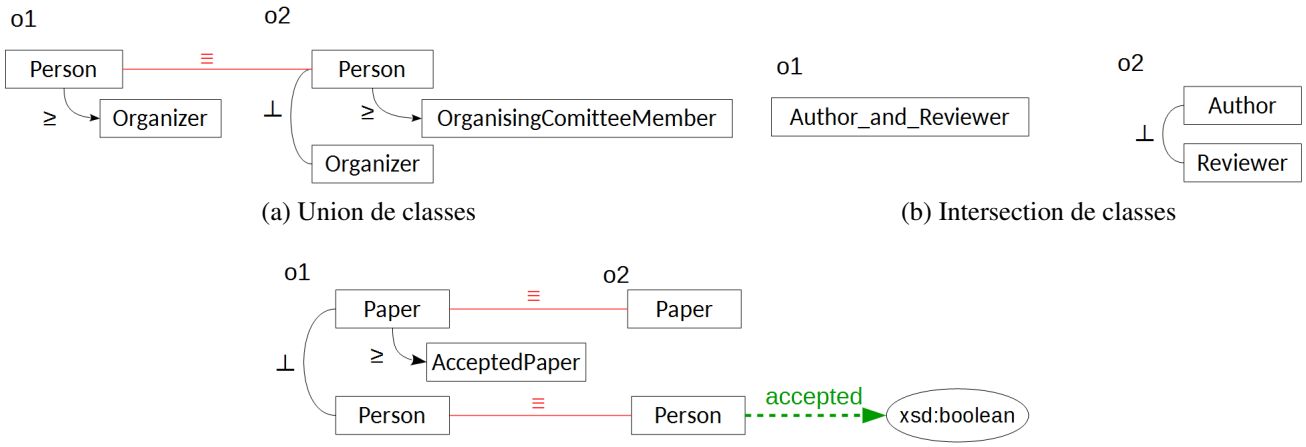
Pour l'étape 3, nous proposons des conditions de cohérence à partir des patrons unitaires de Scharffe (2009). En effet, les patrons permettent de décomposer les correspondances complexes, de les classifier, donc de proposer des conditions de cohérence génériques. Pour assurer la cohérence d'un alignement, il faut s'assurer de la cohérence de toutes ses correspondances individuellement et dans leur ensemble. Individuellement, aucun membre d'une correspondance ne doit représenter un ensemble vide (figure 3b, contraintes de disjonction dans $o2$ seulement : $\forall x$, *o1#Author*(x) $\perp$ *o2#Reviewer*(x)). Considérées dans leur ensemble, les correspondances ne doivent pas se contredire. La disjonction entre deux classes peut venir d'une disjonction couplée à une équivalence : dans la figure 2 la correspondance $\forall x$, *cmt#Reviewer*(x) $\equiv$ *conference#Reviewer*(x) (1) contredit $\forall x, y$, *cmt#assignExternalReviewer*(x, y) $\equiv$ *conference#invites_co-reviewer*(x, y) (2) :

— $(1) \Rightarrow \forall x$, *cmt#ExternalReviewer*$(x)$ $\perp$ *conference#Reviewer*(x)
— $(2) \Rightarrow \forall x$, *cmt#ExternalReviewer*$(x)$ $\equiv$ *conference#Reviewer*(x)

Ces deux correspondances ne peuvent donc pas appartenir au même alignement.

3.2 Correspondances complexes entre classes

Pour chaque classe $c1$ de $o1$ alignée, on vérifie les conditions de cohérence suivantes présentées pour des patrons unitaires. Elles sont toutes à considérer pour une composition de patrons.



(a) Union de classes

(b) Intersection de classes

(c) Restriction de classes

FIGURE 3 – Exemple de patrons de correspondance de classes

Union de classes Pour garantir la cohérence d'une union de classes $c2_i$ avec $c1$, il faut que $c1$ ne soit disjointe avec aucune des $c2_i$. Dans le cas présenté dans la figure 3a, la correspondance $\forall x$, *o1#Organizer*$(x) \geq$ *o2#OrganisingCommitteeMember*$(x) \vee$ *o2#Organizer*(x) n'est pas cohérente car *o1#Organizer* et *o2#Organizer* sont disjointes par extension de la correspondance simple et de la disjonction dans $o2$.

Intersection de classes Dans le cas où $c1$ est mise en correspondance avec une conjonction de classes $c2_i$, $c1$ ne doit pas être disjointe avec aucune des $c2_i$. Les $c2_i$ ne doivent pas non plus être disjointes entre elles. Dans le cas présenté dans la figure 3b, la correspondance $\forall x$, *o1#Author_and_Reviewer*$(x) \equiv$ *o2#Author*$(x) \wedge$ *o2#Reviewer*(x) n'est pas cohérente car *o2#Reviewer* et *o2#Author* sont disjointes, leur intersection vaut donc l'ensemble vide.

Restriction de classes Ces conditions de cohérence sont valables pour les patrons appelés *restrictions de classe* : CAV, CAT, restrictions d'occurrence d'un attribut (CAO). Si $c1$ est mise en correspondance avec une restriction de classe basée sur un attribut (relation ou propriété) $r2$, $c1$ doit être subsumée par le domaine de $r2$, par conséquent, $c1$ ne doit pas être disjointe avec le domaine de $r2$. Dans la figure 3c, la correspondance $\forall x$, *o1#AcceptedPaper*$(x) \equiv$ *o2#Paper*$(x) \wedge$ *o2#accepted*$(x$*,true*$)$ n'est pas cohérente. *o1#Paper* et *o2#Person* sont disjointes par extension des correspondances simples et de la disjonction dans $o1$.

3.3 Correspondances complexes entre relations

Pour chaque relation $r1$ de domaine $c1d$ et de co-domaine $c1r$ dans les correspondances 1:n produites, on vérifie les conditions de cohérence suivantes. Si $r1$ est mis en équivalence avec une construction de relation(s) $r2$, le domaine $c2d$ résultant de $r2$ doit être équivalent à $c1d$ et le co-domaine $c2r$ résultant de $r2$ doit être équivalent à $c2r$. Dans le cas d'une subsomption $\forall x, y, r1(x, y) \leq r2(x, y)$, il faut $c1d \subseteq c2d$ et $c1r \subseteq c2r$. Les domaines et co-domaines résultant d'une construction $r2$ suivant le type de patron mis en oeuvre sont les suivants :

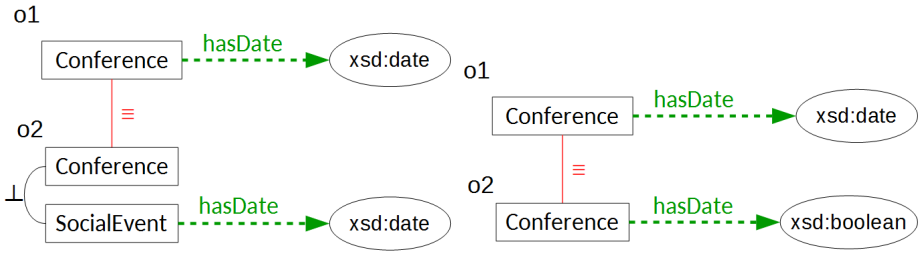
Union de relations $r2$ est une union de relations $r2_i$. $c2d$ vaut l'union des domaines des $r2_i$. $c2r$ vaut l'union des co-domaines des $r2_i$.

Intersection de relations $r2$ est une conjonction de relations $r2_i$. $c2d$ vaut l'intersection des domaines des $r2_i$. $c2r$ vaut l'intersection des co-domaines des $r2_i$.

Inverse d'une relation $r2$ est l'inverse d'une relation $r2_0$. $c2d$ vaut le co-domaine de $r2_0$. $c2r$ vaut le domaine de $r2_0$.

Chaîne de relations $r2$ est une chaîne de relations $r2_i = \{r2_0, ..., r2_n\}$. $c2d$ vaut le domaine de la première relation de la chaîne : $r2_0$. $c2r$ vaut le co-domaine de la dernière relation de la chaîne : $r2_n$. Pour la cohérence d'une chaîne, il faut aussi s'assurer que le co-domaine d'une relation $r2_{j-1}$ ne soit pas disjoint avec le domaine de la relation $r2_j$ suivante dans la chaîne.

3.4 Correspondances complexes entre propriétés sur les données



(a) Incohérence liée au domaine (b) Incohérence de type de données

FIGURE 4 – Fragments d'ontologies ayant des propriétés sur les données

Pour chaque relation $p1$ de domaine $c1$ et de type de données $t1$ mise en correspondance dans l'alignement, on vérifie les conditions de cohérence suivantes. Comme pour les conditions de cohérence sur les relations, si $p1$ est mis en équivalence avec une construction de propriété(s) $p2$, le domaine $c2$ résultant de $p2$ doit être équivalent à $c1$ et le type de données $t2$ résultant de $p2$ doit être compatible avec $t1$. Nous nous basons sur la définition de la compatibilité employée dans Ritze *et al.* (2009) : deux types de données sont compatibles si on peut traduire l'un en l'autre. Dans le cas d'une subsomption $\forall x, y, p1(x, y) \leq p2(x, y)$, il faut $c1 \subseteq c2$ et $t1$ compatible avec $t2$. Pour les patrons d'union, d'intersection et de chaîne, le domaine résultant de la construction $p2$ est le même que pour des patrons de relations. Dans la figure 4a, $\forall x, y$, *o1#hasDate*(x, y) $\equiv$ *o2#hasDate*(x, y) est incohérente car les domaines de ces deux propriétés ne sont pas équivalents (et même disjoints). Dans la figure 4b, $\forall x, y$, *o1#hasDate*(x, y) $\equiv$ *o2#hasDate*(x, y) est incohérente car les types de données sont incompatibles. On distingue deux cas de figure pour vérifier la compatibilité des types de données :

Patrons sans transformation Si $p2$ est une construction impliquant des propriétés $p2_i$ sans fonction de transformation, chaque type de données des $p2_i$ doit être compatible avec $t1$. Cette condition est valable pour les unions, intersections, chaînes de propriétés.

Fonctions de transformation Le type de données résultant d'une fonction de transformation appliquée à des propriétés dépend de cette fonction. Il doit être compatible avec $t1$.

3.5 Filtrage des correspondances

Les correspondances sont ensuite filtrées globalement pour éviter la redondance. Si un alignement simple déclare deux entités $e1$ et $e2$ (de $o1$ et $o2$ respectivement) équivalentes, les

correspondances complexes 1:n ayant $e1$ pour entité source sont supprimées. Par exemple, $\forall x$, *o1#Reviewer(*x*)* $\equiv$ *o2#Reviewer(*x*)* et $\forall x$, *o1#Reviewer(*x*)* $\equiv \exists y$, *o2#reviewes(*x, y*)* sont deux correspondances ayant pour entité source *o1#Reviewer*. On garde la première correspondance et supprime la seconde. Les correspondances dressant une équivalence sont préférées à celle dressant une subsomption (i.e. une correspondance complexe d'équivalence est préférée à une correspondance simple de subsomption) car on peut déduire des subsomptions à partir d'équivalences mais pas l'inverse. Les constructions dans une correspondance complexe sont exprimées de la manière la plus simple possible : le moins d'entités et de constructeurs possibles. Par exemple, on utilisera dans une construction une relation dans son sens direct et non le constructeur inverse appliqué à la relation inverse si cela est possible.

4 Le jeu de données de correspondances complexes

Le jeu de données conférence est utilisé comme banc de test dans l'OAEI (Achichi *et al.* (2016)). Ce jeu de données se compose de 16 ontologies sur le domaine de l'organisation de conférences et d'alignements simples de référence entre 7 de ces ontologies. Ce jeu de données a été choisi car il est fréquemment utilisé dans le domaine de l'alignement d'ontologies (Zamazal & Svátek, 2017). Les ontologies du jeu de données conférence comportent des axiomes, peu d'annotations et présentent plus de réalisme que le jeu de données benchmark synthétique par exemple. Parmi les ontologies du jeu de données conférence, trois ontologies formant deux paires ont été utilisées pour créer le jeu de données de correspondances complexes : *cmt-conference* et *cmt-edas*. Leurs alignements de référence enrichis manuellement ont été utilisés dans la méthodologie. Les alignements complexes du jeu de données sont disponibles en ligne [3].

	cmt	conference	edas
Nombre de classes	30	60	104
Nombre de relations	49	46	30
Nombre de propriétés	10	18	20

TABLE 1 – Caractéristiques des trois ontologies du jeu de données.

La table 1 présente le nombre de classes, relations et propriétés des ontologies. La table 2 présente le type (patron) des correspondances dans le jeu de données par paire d'ontologies et par type de l'entité traduite. L'ontologie source est écrite en premier dans la paire (dans *cmt-conference*, *cmt* est l'ontologie source et *conference* l'ontologie cible). Les relations des correspondances ($\equiv$, $\geq$, $\leq$) ne sont pas représentées. La table 3 montre le nombre d'entités de chaque ontologie source traduites par équivalence dans l'ontologie cible. Chaque case a pour forme $(n_s + n_c)/n_t$ où n_s est le nombre d'entités traduites par une équivalence simple, n_c le nombre d'entités traduites par une équivalence complexe et n_t le nombre total d'entité de ce type dans l'ontologie source. La table 4 présente des exemples de correspondances et leur type.

Évaluation du jeu de données

Il n'existe pas encore de méthode automatique permettant d'évaluer la qualité d'alignements

3. `https://cloud.irit.fr/index.php/s/JMTMonRBadOzzM2`
`https://cloud.irit.fr/index.php/s/gJdcRj0PT5fv4Fd`

complexes. Toutefois, en s'inspirant de Meilicke & Stuckenschmidt (2008), les ontologies alignées ont été fusionnées à partir des correspondances obtenues dans l'alignement. La consistance de l'ontologie ainsi formée a été vérifiée par le raisonneur HermiT sous Protégé. Cela a aussi permis d'identifier des correspondances incohérentes dans l'alignement de référence de correspondances simples (fig. 2). Ces correspondances ont été retirées de notre jeu de données.

Paire	Classes	Relations	Propriétés
cmt-conference	13 simples, 2 CAT, 1 union	2 dom, 2 range, 2 chaines, 4 dom-range	1 simple,1 union (with string concatenation)
conference-cmt	13 simples, 2 CAT, 2 CAE (composites)	2 unions	1 simple, 1 dom, 2 string split
cmt-edas	9 simples, 1 CAE, 1 union	5 simples, 1 dom, 2 dom-range	1 union (with string concatenation), 1 dom
edas-cmt	9 simples, 2 CAT	5 simples, 4 unions	1 dom, 2 string split

TABLE 2 – Types de correspondances. *dom* représente une restriction de domaine, *range* représente une restriction de co-domaine, *dom-range* une combinaison des deux.

	Classes traduites	Relations traduites	Propriétés traduites
cmt-conference	(13 + 2)/30	(0 + 10)/49	(1 + 1)/10
conference-cmt	(13 + 4)/60	(0 + 0)/46	(1 + 2)/18
cmt-edas	(9 + 1)/30	(5 + 3)/49	(0 + 2)/10
edas-cmt	(9 + 2)/104	(5 + 0)/30	(0 + 3)/20

TABLE 3 – Entités traduites par équivalence (simple + complexe)/nombre d'entités (classes, relations ou propriétés) de l'ontologie $o1$ (alignement $o1$-$o2$)

entité source	rel.	construction cible	type
$\forall$ x, cmt#ConferenceMember(x)	$\equiv$	$\exists$ y, edas#isMemberOf(x,y)	CAT
$\forall$ x, cmt#Chairman(x)	$\geq$	edas#ConferenceChair(x) $\vee$ edas#SessionChair(x)	union
$\forall$ x, edas#AcceptedPaper(x)	$\equiv$	$\exists$ y, cmt#hasDecision(x,y) $\wedge$ cmt#Acceptance(y)	CAT
$\forall$ x, edas#RejectedPaper(x)	$\equiv$	$\exists$ y, cmt#hasDecision(x,y) $\wedge$ cmt#Rejection(y)	CAT
$\forall$ x, conference#Submitted_contribution(x)	$\equiv$	$\exists$ y, cmt#submitPaper(y,x)	CAE (inverse)
$\forall$ x, conference#Reviewed_contribution(x)	$\equiv$	$\exists$ y, cmt#readByReviewer(x,y) $\vee$ cmt#hasDecision(x,y)	CAE (union)
$\forall$ x,y, cmt#email(x,y)	$\equiv$	edas#Person(x) $\wedge$ edas#hasEmail(x,y)	dom
$\forall$ x,y, edas#hasRelatedDocument(x,y)	$\geq$	cmt#writeReview(x,y) $\vee$ cmt#writePaper(x,y)	union
$\forall$ x,y cmt#assignedTo(x,y)	$\equiv$	$\exists$ z, conference#has_a_review(x,z) $\wedge$ conference#has_author(z,y)	chaine relation

TABLE 4 – Quelques correspondances complexes du jeu de données

5 Evaluation d'approches existantes

Parmi les systèmes publiquement disponibles, ceux de Ritze *et al.* (2009, 2010) ne nécessitent pas d'instances communes. Peupler les deux ontologies avec des instances communes

ajoute de nouvelles considérations à prendre en compte pour la création du jeu de données car le choix et la quantité des instances ne sont pas anodins. Pour cette raison, seules les deux approches de Ritze *et al.* sont évaluées sur le jeu de données que nous proposons. Ces approches prennent pour input l'alignement simple de référence et les deux ontologies à aligner.

5.1 Evaluation de Ritze *et al.* (2009)

La première approche proposée par Ritze *et al.* (2009) trouve les correspondances suivantes entre *cmt* et *conference* :

— $\forall x$, *cmt#AuthorNotReviewer(*x*)* $\equiv$ $\exists y$, *conference#contributes(*x, y*)* $\wedge$ *conference#Reviewed_contribution(*y*)*
(Un auteur non-relecteur est une personne qui contribue à un article relu)

— $\forall x$, *cmt#Reviewer(*x*)* $\equiv$ $\exists y$, *conference#contributes(*x, y*)* $\wedge$ *conference#Reviewed_contribution(*y*)*
(Un relecteur est une personne qui contribue à un article relu)

Ces deux correspondances, sans doute obtenues par similarité des chaînes de caractères "Reviewer" et "Reviewed", sont fausses. Le rappel et la précision sont 0 pour *cmt-conference*.

Entre *cmt* et *edas*, les correspondances suivantes sont détectées :

— $\forall x$, *edas#RejectedPaper(*x*)* $\equiv \exists y$, *cmt#hasDecision(*x, y*)* $\wedge$ *cmt#Rejection(*y*)*
(Un papier rejeté est un papier ayant une décision de rejet)

— $\forall x$, *edas#AcceptedPaper(*x*)* $\equiv \exists y$, *cmt#hasDecision(*x, y*)* $\wedge$ *cmt#Acceptance(*y*)*
(Un papier accepté est un papier ayant une décision d'acceptation)

— $\forall x$, *cmt#ConferenceMember(*x*)* $\equiv \exists y$, *edas#hasMember(y,x)* $\wedge$ *edas#Person(*x*)*
(Un membre d'une conférence est une personne qui a été membre d'une conférence)

— $\forall x$, *cmt#AuthorNotReviewer(*x*)* $\equiv \exists y$, *edas#hasRelatedDocument(*x, y*)* $\wedge$ *edas#Review(*y*)*
(Un auteur non-relecteur est une personne qui contribue à une relecture.)

Les trois premières correspondances sont correctes et figurent dans l'alignement *cmt-edas* (resp. *edas-cmt*) proposé. La correspondance $\forall x$, *cmt#ConferenceMember(*x*)* $\equiv$ $\exists y$, *edas#hasMember(y,x)* $\wedge$ *edas#Person(*x*)* n'est pas écrite sous la même forme mais une correspondance équivalente y est : $\forall x$, *cmt#ConferenceMember(*x*)* $\equiv$ $\exists y$, *edas#isMemberOf(*x, y*)*. Cette correspondance est équivalente car *edas#isMemberOf* est la relation inverse de *edas#hasMember*. La précision est 0.75 (3/4) et le rappel (sur l'ensemble des correspondances complexes par équivalence dans les deux sens (*cmt-edas* et *edas-cmt*)) 0.27 (3/11).

5.2 Evaluation de Ritze *et al.* (2010)

L'approche de Ritze *et al.* (2010) à partir de l'alignement de référence n'a pas permis de détecter de correspondances complexes entre *cmt* et *conference*. Même si des correspondances de type CAT sont présentes entre cmt et conference (e.g. $\forall x$, *cmt#ProgramCommitteeMember(*x*)* $\equiv$ $\exists y$, *conference#was_a_member_of(*x, y*)* $\wedge$ *conference#Program_committee(*y*)*), elles sont difficiles à détecter par l'approche de Ritze *et al.* (2010) car les conditions de correspondance sont très restrictives. En effet, le *modificateur* (*ProgramCommittee*) du *substantif principal* (*Member*) n'est pas clairement délimité car *ProgramCommittee* est lui-même un nom composé. D'autre part, le modificateur *ProgramCommittee* n'est pas détecté comme nominalisation de *Program_committee*. Dans l'alignement *conference-cmt*, deux autres CAT auraient dû être

détectés. Toutefois, les restrictions linguistiques trop strictes n'ont pas permis de les identifier. Entre *cmt* et *edas*, Ritze *et al.* (2010) détecte les correspondances suivantes :

— $\forall x$, *edas#RejectedPaper(*x*)* $\equiv \exists y$, *cmt#hasDecision(*x, y*)* $\wedge$ *cmt#Rejection(*y*)*
(Un papier rejeté est un papier ayant une décision de rejet)

— $\forall x$, *edas#AcceptedPaper(*x*)* $\equiv \exists y$, *cmt#hasDecision(*x, y*)* $\wedge$ *cmt#Acceptance(*y*)*
(Un papier accepté est un papier ayant une décision d'acceptation)

Il n'y a pas de correspondance incorrecte. Toutefois, la correspondance $\forall x$, *cmt#ConferenceMember(*x*)* $\equiv$ $\exists y$, *edas#hasMember(y,x)* $\wedge$ *edas#Person(*x*)* n'est plus détectée. La précision est donc de 1 (2/2) et le rappel (sur l'ensemble des correspondances complexes par équivalence dans les deux sens (*cmt-edas* et *edas-cmt*)) de 0.18 (2/11). Les apports linguistiques de cette proposition améliorent la précision de l'approche précédente. Cependant, ils empêchent la détection de certaines correspondances et diminuent le rappel.

6 Discussion

L'analyse du jeu de données permet d'en observer les limites et mettre en évidence ses améliorations futures. Son utilisation pour évaluer des approches d'alignement complexe nous a également permis de mettre en exergue leurs limites. Les approches de Ritze *et al.* contiennent peu de conditions sur les correspondances entre relations et entre propriétés. De plus ces approches ne détectent pas des patrons composés. Si les ontologies étaient peuplées avec des instances communes, l'approche de Nunes *et al.* (2011) aurait pu détecter des combinaisons de propriétés. De même, Qin *et al.* (2007) aurait pu découvrir les chaînes de relations et/ou de propriétés présentes. Les approches de Walshe (2014) et Parundekar *et al.* (2010, 2012) ne découvriraient aucune correspondance du jeu de données car il ne comporte pas de CAV. Parmi les approches d'alignement complexe entre ontologies présentées ici, aucune ne serait en mesure de détecter des correspondances composées (patrons composés). Le jeu de données met en évidence quelques pistes d'amélioration des approches d'alignement complexe. Les axiomes et définitions sont source d'information mais sont peu (ou pas) utilisés par les approches existantes. Les bases de patrons des approches les utilisant peuvent être élargies. Certaines conditions de détection de correspondance par patron (de Ritze *et al.* (2010)) pourraient être complétées ou relâchées. Le jeu de données comporte certains patrons unitaires (CAT, restriction de domaine d'une relation, etc.), d'autres composés. Toutefois, certains patrons de correspondance ne sont pas représentés dans le jeu de données. En effet, aucun CAV n'est présent, de même pour les CAO (restriction de classe par occurrence d'un attribut). Même si on dénombre quelques CAE (restriction de classe par existence d'un attribut), qui sont des cas spéciaux de CAO, ils ne sont pas unitaires dans ce jeu de données mais composés. Aucune intersection n'est présente. Le jeu de données pourrait donc être étendu pour y ajouter d'autres paires d'ontologies dont les correspondances contiendraient des patrons pour l'instant absents. Une autre piste à explorer est le peuplement des ontologies pour permettre aux approches nécessitant des instances d'être évaluées sur ce jeu. Les alignements proposés pourraient aussi être exprimés en EDOAL, une syntaxe dédiée, intégrée à l'Alignment API, pour faciliter son intégration aux outils existants d'évaluation. D'autre part, certains choix de correspondance ont été faits lors de la création du jeu de données (section 3.5). La pertinence de ces choix pourrait être validée par la comparaison des correspondances établies par plusieurs experts qui suivraient la méthode proposée. Un dernier point d'amélioration serait d'ajouter des correspondances n:m au jeu de données.

7 Conclusion

Les alignements complexes sont nécessaires pour pallier l'hétérogénéité entre ontologies. Ils complètent les alignements simples en exprimant plus finement les niveaux de disparités. L'évaluation des approches d'alignement complexe est un besoin croissant. Pour l'instant, leur évaluation est basée sur la précision. Le jeu de données proposé ici est exhaustif pour les correspondances d'équivalence 1:n entre 2 paires d'ontologies issues du jeu de données conférence de l'OAEI. L'évaluation d'approches d'alignement complexe sur le jeu de données permet de mettre en avant leurs limites. Le jeu de données peut être étendu pour servir à l'évaluation de approches utilisant et nécessitant des instances. Les correspondances n:m sont aussi un autre aspect à développer. Il serait intéressant, comme dans Cheatham & Hitzler (2014) de faire valider ce jeu de données par plusieurs experts ainsi que d'explorer la confiance des correspondances.

Références

ACHICHI M., CHEATHAM M., DRAGISIC Z., EUZENAT J., FARIA D., FERRARA A., FLOURIS G., FUNDULAKI I., HARROW I., IVANOVA V. & OTHERS (2016). Results of the Ontology Alignment Evaluation Initiative 2016. In *11th ISWC workshop on ontology matching (OM)*, p. 73–129.

CHEATHAM M. & HITZLER P. (2014). Conference v2. 0 : An uncertain version of the OAEI Conference benchmark. In *International Semantic Web Conference*, p. 33–48 : Springer.

EUZENAT J. & LE DUC C. (2012). Methodological guidelines for matching ontologies. In *Ontology engineering in a networked world*, p. 257–278. Springer.

EUZENAT J. & SHVAIKO P. (2013). *Ontology Matching*. Springer Berlin Heidelberg.

KLEIN M. (2001). Combining and relating ontologies : an analysis of problems and solutions. In *IJCAI-2001 Workshop on ontologies and information sharing*, p. 53–62 : USA.

MEILICKE C. & STUCKENSCHMIDT H. (2008). Incoherence as a basis for measuring the quality of ontology mappings. In *3rd ISWC workshop on ontology matching (OM)*, p. 1–12.

NUNES B. P., MERA A., CASANOVA M. A., BREITMAN K. K. & LEME L. A. P. (2011). Complex Matching of RDF Datatype Properties. In *6th ISWC workshop on ontology matching (OM)*.

PARUNDEKAR R., KNOBLOCK C. A. & AMBITE J. L. (2010). Linking and building ontologies of linked data. In *International Semantic Web Conference*, p. 598–614 : Springer.

PARUNDEKAR R., KNOBLOCK C. A. & AMBITE J. L. (2012). Discovering concept coverings in ontologies of linked data sources. In *International Semantic Web Conference*, p. 427–443 : Springer.

QIN H., DOU D. & LEPENDU P. (2007). Discovering executable semantic mappings between ontologies. In *On the Move to Meaningful Internet Systems*, p. 832–849 : Springer.

RITZE D., MEILICKE C., ŠVÁB ZAMAZAL O. & STUCKENSCHMIDT H. (2009). A pattern-based ontology matching approach for detecting complex correspondences. In *4th ISWC workshop on ontology matching (OM)*, p. 25–36.

RITZE D., VÖLKER J., MEILICKE C. & ŠVÁB ZAMAZAL O. (2010). Linguistic analysis for complex ontology matching. In *5th ISWC workshop on ontology matching (OM)*, p. 1–12.

SCHARFFE F. (2009). *Correspondence Patterns Representation*. PhD thesis, Faculty of Mathematics, Computer Science and University of Innsbruck.

WALSHE B. (2014). *Detecting Restriction Class Correspondences in Linked Open Data*. PhD thesis, Trinity College, Dublin.

ZAMAZAL O. & SVÁTEK V. (2017). The Ten-Year OntoFarm and its Fertilization within the Onto-Sphere. *Web Semantics : Science, Services and Agents on the World Wide Web*, **43**, 46–53.

ADEL : une méthode adaptative de désambiguïsation d'entités nommées

Julien Plu[1], Raphaël Troncy[1], Giuseppe Rizzo[2]

[1] EURECOM, Sophia-Antipolis, France
julien.plu@eurecom.fr, raphael.troncy@eurecom.fr

[2] ISMB, Turin, Italie giuseppe.rizzo@ismb.it

Résumé : Nous identifions quatre critères principaux pouvant causer de sérieuses difficultés lorsqu'il s'agit de développer un système de désambiguïsation d'entités nommées : i) la nature du document à analyser (tweet, sous-titre d'une vidéo, article de presse) ; ii) le nombre de types possibles qui sont utilisés pour catégoriser une mention (Personne, Lieu, Date, Rôle) ; iii) la base de connaissances utilisée pour désambiguïser les mentions extraites (DBpedia, Wikidata, Musicbrainz) ; iv) la langue utilisée dans le document. Dans cet article, nous présentons ADEL, une approche innovante basée sur une architecture hybride pour un système de désambiguïsation d'entités nommées combinant des méthodes du domaine du Traitement Automatique du Langage Naturel (TALN), de la recherche d'information et du Web Sémantique. En particulier, nous proposons une approche modulaire afin d'être aussi indépendants que possible du texte analysé et de la base de connaissance utilisée. Notre évaluation montre que ce système obtient des résultats comparables ou meilleurs que l'état de l'art sur cinq jeux de données : OKE2015, OKE2016, NEEL2014, NEEL2015 et NEEL2016.

Mots-clés : Désambiguïsation d'entités nommées, reconnaissance d'entités nommées, extraction d'information

1 Introduction

Il y a une augmentation exponentielle de contenu textuel disponible sur le Web, notamment produit par les internautes via une large diversité de plate-forme de publication et qui a besoin d'être traité automatiquement. Concernant ce contenu textuel, nous avons identifié quatre défis principaux pour la communauté du TALN : gérer différents types de textes (billets sur les réseaux sociaux, requêtes d'un moteur de recherche, sous-titres d'une vidéo, articles de journaux) écrits dans des langues différentes ; détecter des entités nommées propres à des domaines variés pouvant être classifiées et désambiguïsées avec des classes et des graphes de connaissances spécifiques. Ces défis affectent la stratégie à utiliser pour comprendre le texte et pour extraire et désambiguïser des unités d'informations pertinentes. Différentes bases de connaissances ont été utilisées pour une telle tâche : DBpedia, Freebase, Wikidata [1] pour n'en nommer que quelques-unes. Ces bases de connaissances sont connues pour être large en termes de couverture, alors que des bases de connaissances verticales existent aussi dans des domaines plus spécifiques, par exemple, Geonames, Musicbrainz ou LinkedMDB [2]. Ces quatre défis sont bien représentés dans différents jeux de donnée de références (van Erp *et al.*, 2016), où chaque jeu de données a ses propres règles d'annotation : différent contenu textuel (articles de journaux pour AIDA, tweets pour NEEL), différentes définitions des types des entités nommées (Location pour NEEL, Place pour OKE), différentes base de connaissances (Freebase pour TAC-KBP, DBpedia pour OKE), différentes langues (anglais pour MEANTIME, français pour ETAPE (Gravier *et al.*, 2012)).

1. http://dbpedia.org, https://www.freebase.com, https://www.wikidata.org/
2. http://www.geonames.org, https://musicbrainz.org, http://www.linkedmdb.org

La désambiguïsation et la reconnaissance d'entités nommées sont deux sous-tâches de l'extraction d'information. Dans cet article, nous appelons une *mention* - une expression extraite d'un texte, une *entité nommée* - une ressource décrite dans une base de connaissances, une *entité nommée candidate* - une possible entité nommée pour une mention, et une *annotation* comme le couple *(mention, entité nommée)* définissant le lien établi d'une mention dans le texte à la base de connaissance référente. Le principe de la reconnaissance d'entités nommées est d'extraire des mentions et de leur donner un type conformément à un ensemble de classes prédéfinies (par exemple Personne, Lieu ou Organisation). Le principe de désambiguïsation d'entités nommées est d'annoter les mentions extraites à partir du texte avec leur entité correspondante décrite dans une base de connaissances. Chaque entité nommée se voit associer une liste de candidats. Toutes les entrées dans la base de connaissance représentent une entité du monde réel de manière unique avec un identifiant spécifique. Les entités nommées qui n'apparaissent pas dans la base de connaissance sont généralement appelées *entités nommées émergentes* ou *NIL*. Lorsque l'on analyse du contenu textuel, l'ambiguïté et la synonymie sont des problèmes fondamentaux dont il faut s'occuper. Une entité nommée peut avoir plus d'une mention (synonymie) et une mention peut représenter plus d'une entité nommée (ambiguïté). Dans cet article, nous proposons une approche de reconnaissance et désambiguïsation d'entités nommées permettant de répondre à ces quatre défis. Nous détaillons notre approche et le système ADEL dans la Section 2 et nous décrivons plusieurs évaluations dans la Section 3 avant de conclure (Section 4).

2 ADEL : une architecture modulaire

ADEL est composé de multiples modules qui peuvent être activés ou désactivés. Ces modules sont regroupés dans trois composants : *Entity Extraction*, *Index* et *Entity Linking* (Figure 2). ADEL est implémenté en Java et est publiquement disponible comme une API REST [3]. ADEL s'attaque aux défis mentionnés dans l'introduction en adaptant ses fonctionnalités au texte, à la langue, au type des entités nommées et à la base de connaissance.

2.1 Entity Extraction

Le premier module du composant *Entity Extraction* est le *Extractors Module*. Il extrait les mentions à partir du texte qui sont susceptibles d'être sélectionnées comme entités nommées. Après avoir identifié les mentions, nous résolvons leur potentielle superposition en utilisant le module *Overlap Resolution Module*.

Nous utilisons de multiples extracteurs : dictionnaire, étiquetage morphosyntaxique (POS), reconnaissance d'entités nommées (NER), co-référence, nombre et date. Chacun de ces extracteurs fonctionne en parallèle et peut être basé sur des systèmes de TALN extérieurs comme Stanford CoreNLP, GATE ou NLTK. Nous avons développé une API Web générique pouvant être implémentée pour chaque système de TALN afin de le rendre utilisable par ADEL. Comme exemple, la version de cette API pour Stanford CoreNLP est disponible sur Github [4]. Nous fournissons aussi une documentation de cette API [5] afin de permettre à tout le monde de développer

3. `http://adel.eurecom.fr/api/`
4. `https://github.com/jplu/stanfordNLPRESTAPI`
5. `https://github.com/jplu/stanfordNLPRESTAPI/wiki/API-documentation`

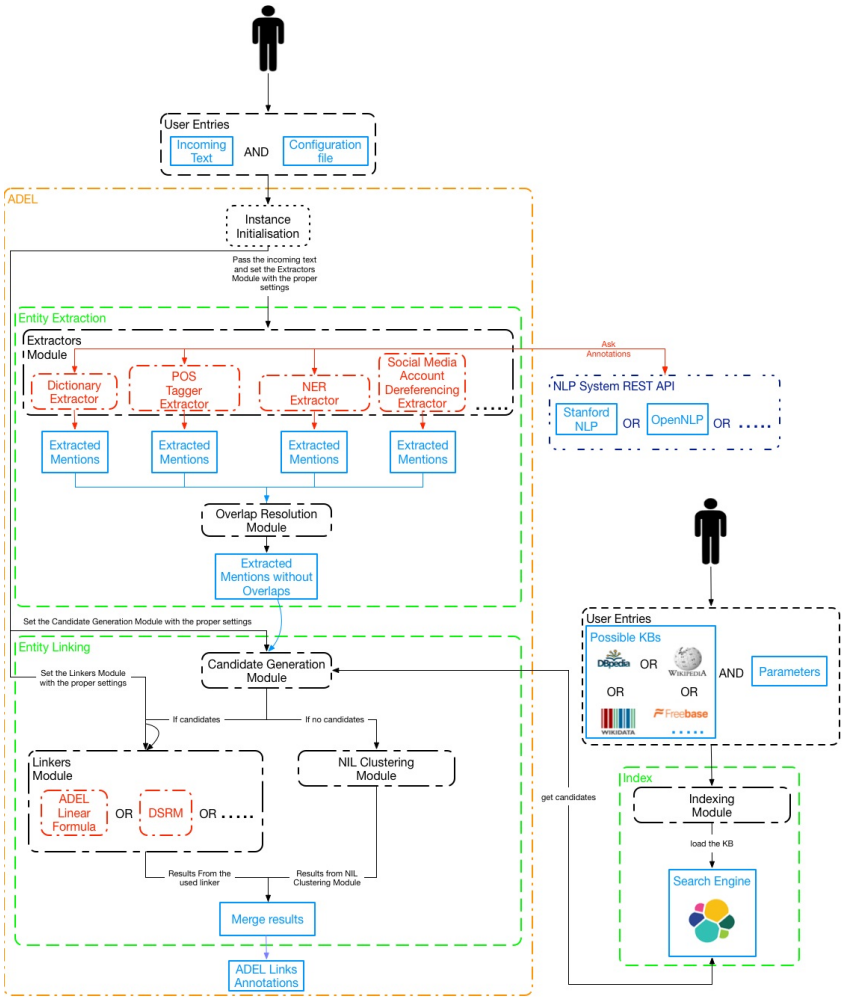


FIGURE 1 – Architecture du système ADEL

son propre extracteur à intégrer dans ADEL. Il n'y a pas de limite au nombre d'extracteurs et de modèles pouvant être utilisé dans ADEL. L'API Web de chaque extracteur est utilisé comme composant extérieur, et a besoin d'être instancié en amont. Il est possible d'utiliser une combinaison de modèles permettant de faire usage de différents modèles CRF. Soit f une fonction qui prend un modèle m et un texte txt et produit un ensemble de tuples $(mention, etiquette)$: $f_m(txt) = \{(c_1, l_1), \ldots, (c_k, l_k)\}$. La combinaison des modèles revient à faire l'union des résultats obtenus pour chaque modèle considéré : $\bigcup_{m \in M} f_m(txt)$. L'ordre dans lequel les modèles sont appliqués est important. Par conséquent, si une mention est incorrectement annotée par le premier modèle, le second modèle ne pourra pas le corriger même si ce second modèle aurait pu annoter correctement cette mention. Cet algorithme dans Stanford NER est appelé *NER Classifier Combiner*.

Un extracteur peut extraire des mentions qui ont une superposition partielle ou totale avec les autres mentions extraites par d'autres extracteurs. Afin de résoudre cette ambiguïté, nous avons implémenté un module de résolution de superposition qui prend la sortie de chaque extracteur et donne une sortie sans superpositions. La logique de ce module est la suivante : étant donné deux

mentions qui se superposent, par exemple `Unis d'Amérique` venant de l'extracteur NER et `Etats-Unis` venant de l'extracteur POS, nous prenons l'union des deux mentions. Nous obtenons la mention `Etats-Unis d'Amérique` et le type fourni par l'extracteur NER est sélectionné. Nous avons aussi implémenté d'autres heuristiques afin de résoudre les superpositions, mais le choix de la bonne heuristique à utiliser est laissé à une configuration manuelle.

2.2 Index

Des bases de connaissances différentes sont utilisées en fonction des *challenges* de référence (par exemple, Wikipedia et ensuite Freebase comme base de connaissances pour TAC-KBP). Il y a donc un besoin d'avoir une façon générique et performante d'indexer n'importe quelle base de connaissances. Un index peut être vu comme un tableau à deux dimensions où chaque ligne est une entité dans l'index et chaque colonne est une propriété qui décrit par un littéral cette entité. Ainsi, indexer entièrement la version anglaise de DBpedia donne 281 colonnes. Une fois que nous avons cet index, nous pouvons chercher une mention dans cet index et récupérer les entités nommées candidates. Chercher, par défaut, sur toutes les colonnes (ou les propriétés utilisées dans la base de connaissances), impact de manière négative la performance de l'index en termes de temps de calcul. Afin d'optimiser l'index, nous avons développé une heuristique qui vise à maximiser la couverture de l'index tout en minimisant le nombre de colonnes (ou propriétés) sur lesquelles effectuer une recherche. Pour la version 2015-10 de DBpedia, il y a exactement 281 propriétés ayant des valeurs littérales, alors que notre méthode d'optimisation réduit cette liste à 8 propriétés : `dbo:wikiPageWikiLinkText`, `dbo:wikiPageRedirects`, `dbo:demonym`, `dbo:wikiPageDisambiguates`, `dbo:birthName`, `dbo:alias`, `dbo:abstract` et `rdfs:label`. Cette optimisation réduit drastiquement le temps de la requête en passant d'environ 4 secondes à moins d'une seconde. Le code source de cette optimisation est aussi disponible [6]. L'index est construit en utilisant *Elasticsearch* comme moteur de recherche. L'indexation d'une base de connaissances suit un processus à deux étapes : *i)* extraire le contenu de la base de connaissances et créé l'index Elasticsearch avec ces données ; *ii)* lancer la méthode d'optimisation afin de récupérer la liste des colonnes qui seront utilisées pour interroger l'index.

2.3 Entity Linking

Le composant *Entity Linking* commence avec le *Candidate Generation Module* qui interroge l'index et génère une liste d'entités nommées candidates pour chaque mention extraite. Si l'index retourne une liste d'entités nommées candidates, alors le *Linkers Module* est invoqué ; alternativement, si une liste vide est retournée, alors le *NIL Clustering Module* est invoqué.

Le *NIL Clustering Module* propose de regrouper les entités nommées *NIL* (entités nommées émergentes) qui peuvent identifier la même chose dans le monde réel. Le rôle de ce module est d'attacher la même valeur *NIL* dans un même document. Par exemple, si deux mentions extraites partagent la même entité nommée émergente, ces mentions seront annotées avec la même valeur *NIL_1*.

Le *Linkers Module* permet d'offrir plusieurs méthodes. Jusque là, nous avons intégré une seule méthode appelée *ADEL Formula*, mais nous sommes actuellement en train de dévelop-

6. `https://gist.github.com/jplu/a16103f655115728cc9dcff1a3a57682`

per d'autres méthodes qui seront aussi intégrées dans ADEL. Ces futures méthodes seront basées sur D2V Le & Mikolov (2014), TF-IDF avec une distance cosinus, ainsi qu'adapter DSSM Huang *et al.* (2013) pour la désambiguïsation d'entités nommées. Nous avons aussi développé une fonctionnalité dans ADEL permettant l'intégration d'une méthode de désambiguïsation extérieure en respectant l'implémentation d'une interface Java. La méthode *ADEL Formula* se résume à l'Equation 1.

$$r(l) = (a \cdot L(m, label) + b \cdot max(L(m, R)) + c \cdot max(L(m, D))) \cdot PR(l) \quad (1)$$

La fonction $r(l)$ utilise la distance de Levenshtein L entre la mention m et le label, la distance maximum entre la mention m et chaque élément (label) dans l'ensemble de pages de redirection de Wikipedia R et la distance maximum entre la mention m et chaque élément (label) dans l'ensemble de pages de désambiguïsation de Wikipedia D, pondéré par le PageRank PR, pour chaque entité nommée candidate l. Les poids a, b et c sont une combinaison convexe devant satisfaire : $a + b + c = 1$ et $a > b > c > 0$. Nous prenons l'hypothèse que la distance entre une mention et un label est plus importante que la distance avec la page de redirection et qui est elle-même plus importante que la distance avec la page de désambiguïsation.

3 Evaluation

Notre approche a été évaluée avec les jeux de données tests de plusieurs *challenges* : NEEL-2014 (Basave *et al.*, 2014), NEEL2015 (Rizzo *et al.*, 2015) et NEEL2016 (Rizzo *et al.*, 2016), ainsi que OKE2015 (Nuzzolese *et al.*, 2015) et OKE2016 (Nuzzolese *et al.*, 2016). Nous présentons une évaluation de chaque composant pour chacun de ces jeux de données en utilisant le *scorer* TAC-KBP (Hachey *et al.*, 2014) : le Tableau 1 montre les performances d'ADEL pour chacun de ses composants et le Tableau 2 montre les meilleurs scores parmi tous les participants de ces *challenges*.

Niveau	**OKE2015**	**OKE2016**	**NEEL2014**	**NEEL2015**	**NEEL2016**
strong_mention_match	71.2	78.2	70.8	71.6	85.5
strong_typed_mention_match	59.8	72.0	-	52.8	61
strong_link_match	48	49.1	46.3	47.9	53.8

TABLE 1 – Résultats d'ADEL avec le *scorer* NELEVAL en F-mesure

Niveau	**OKE2015**	**OKE2016**	**NEEL2014**	**NEEL2015**	**NEEL2016**
strong_mention_match	**71.2**	**78.2**	-	-	**85.5**
strong_typed_mention_match	**59.8**	**72.0**	-	80.7	**61**
strong_link_match	**48**	60.08	70.6	76.2	**53.8**

TABLE 2 – Meilleur score parmi tous les participants des *challenges* en F-mesure. Les scores en gras sont ceux d'ADEL.

Les résultats pour les *challenges* OKE montrent que nous sommes meilleurs dans tous les cas (sauf un). Les résultats pour les *challenges* NEEL montrent que notre approche est robuste pour l'extraction et la classification des entités nommées voir meilleurs que les autres participants pour l'édition 2016. D'une manière générale, les résultats montrent une chute importante au moment de la désambiguïsation, principalement due à une approche non supervisée.

4 Conclusion

Nous avons démontré qu'ADEL permet de s'adapter aux quatre défis suivants : la **langue** (en changeant les modèles et les dictionnaires utilisés par le(s) système de TALN) ; la nature du **texte** (tweets, sous-titre de vidéos, articles de presse) ; les **classes des entités nommées** (des types conventionnels comme Personne, Lieu ou Organisation à des types plus précis comme Sportif ou Chanteur) ; la **base de connaissances** (DBpedia ou MusicBrainz). Comme futur travail, nous prévoyons d'intégrer trois nouvelles méthodes de désambiguïsation : D2V, TF-IDF avec la similarité de cosinus et DSSM. Nous prévoyons aussi d'améliorer le processus d'extraction pour *1)* améliorer la prise en charge des *hashtags* en intégrant une méthode de segmentation performante pour améliorer la génération de candidats pour les tweets ; *2)* implémenter une association entre les types provenant de plusieurs modèles (Rizzo *et al.*, 2014) afin d'améliorer le raffinement des types.

Références

BASAVE A. E. C., RIZZO G., VARGA A., ROWE M., STANKOVIC M. & DADZIE A. (2014). Making Sense of Microposts (#Microposts2014) Named Entity Extraction & Linking Challenge. In *4th Workshop on Making Sense of Microposts*, Seoul, Korea.

GRAVIER G., ADDA G., PAULSSON N., CARRÉ M., GIRAUDEL A. & GALIBERT O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *8th International Conference on Language Resources and Evaluation (LREC)*.

HACHEY B., NOTHMAN J. & RADFORD W. (2014). Cheap and easy entity evaluation. In *52nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

HUANG P.-S., HE X., GAO J., DENG L., ACERO A. & HECK L. (2013). Learning Deep Structured Semantic Models for Web Search Using Clickthrough Data. In *22nd ACM International Conference on Information & Knowledge Management (CIKM)*.

LE Q. V. & MIKOLOV T. (2014). Distributed Representations of Sentences and Documents. In *31st International Conference on Machine Learning (ICML)*.

NUZZOLESE A., GENTILE A., PRESUTTI V., GANGEMI A., GARIGLIOTTI D. & NAVIGLI R. (2015). The 1st Open Knowledge Extraction Challenge. In *12th European Semantic Web Conference (ESWC)*.

NUZZOLESE A., GENTILE A., PRESUTTI V., GANGEMI A., MEUSEL R. & PAULHEIM H. (2016). The 2nd Open Knowledge Extraction Challenge. In *13th European Semantic Web Conference (ESWC)*.

RIZZO G., BASAVE A. E. C., PEREIRA B. & VARGA A. (2015). Making Sense of Microposts (#Microposts2015) Named Entity rEcognition and Linking (NEEL) Challenge. In *5th Workshop on Making Sense of Microposts*, Florence, Italy.

RIZZO G., VAN ERP M., PLU J. & TRONCY R. (2016). NEEL 2016 : Named Entity rEcognition & Linking challenge report. In *6th International Workshop on Making Sense of Microposts*.

RIZZO G., VAN ERP M. & TRONCY R. (2014). Inductive Entity Typing Alignment. In *2nd International Workshop on Linked Data for Information Extraction (LD4IE)*.

VAN ERP M., MENDES P. N., PAULHEIM H., ILIEVSKI F., PLU J., RIZZO G. & WAITELONIS J. (2016). Evaluating Entity Linking : An Analysis of Current Benchmark Datasets and a Roadmap for Doing a Better Job. In *10th International Conference on Language Resources and Evaluation (LREC)*.

Extraction de relations : combiner les techniques pour s'adapter à la diversité du texte

Adel Ghamnia[1,2], Mouna Kamel[1], Cassia Trojahn[1], Cécile Fabre[2], Nathalie Aussenac-Gilles[1]

[1] IRIT Institut de Recherche en Informatique de Toulouse
Toulouse, France
{adel.ghamnia,mouna.kamel,cassia.trojahn,nathalie.aussenac-gilles}@irit.fr

[2] Laboratoire CLLE, équipe ERSS, Toulouse, France
cecile.fabre@univ-tlse2.fr

Résumé : Extraire des relations d'hyperonymie à partir des textes est une des étapes clés de la construction automatique d'ontologies et du peuplement de bases de connaissances. Plusieurs types de méthodes (linguistiques, statistiques, combinées) ont été exploités par une variété de propositions dans la littérature. Les apports respectifs et la complémentarité de ces méthodes sont cependant encore mal identifiés pour optimiser leur combinaison. Dans cet article, nous nous intéressons à la complémentarité de deux méthodes de nature différente, l'une basée sur les patrons linguistiques, l'autre sur l'apprentissage supervisé, pour identifier la relation d'hyperonymie à travers différents modes d'expression. Nous avons appliqué ces méthodes à un sous-corpus de Wikipedia en français, composé des pages de désambiguïsation. Ce corpus se prête bien à la mise en œuvre des deux approches retenues car ces textes sont particulièrement riches en relations d'hyperonymie, et contiennent à la fois des formulations rédigées et d'autres syntaxiquement pauvres. Nous avons comparé les résultats des deux méthodes prises indépendamment afin d'établir leurs performances respectives, avec le résultat des deux méthodes appliquées ensemble. Les meilleurs résultats obtenus correspondent à ce dernier cas de figure avec une F-mesure de 0.68. De plus, l'extracteur Wikipedia issu de ce travail permet d'enrichir la ressource sémantique DBPedia en français : 55% des relations exprimées et identifiées par notre extracteur ne sont pas présentes dans DBPedia.
Mots-clés : Extraction de relations d'hyperonymie, Supervision distante, Patrons lexico-syntaxiques, Bases de connaissances

1 Introduction

Dans de nombreux domaines tels que l'intelligence artificielle, le web sémantique, le génie logiciel, la recherche d'information ou l'aide au diagnostic, les applications nécessitent un fort potentiel de raisonnement, basé sur une ressource sémantique qui décrit généralement des concepts liés par des relations. Ces ressources peuvent être construites manuellement. Elles sont alors de bonne qualité, mais du fait de leur élaboration coûteuse, elles n'offrent qu'une couverture restreinte du domaine. Compte tenu du volume sans cesse croissant de textes disponibles dans un format numérique, le traitement automatique de la langue permet d'envisager la construction (semi-)automatique de telles ressources, en exploitant les connaissances véhiculées dans ces textes. Un enjeu majeur est alors d'acquérir ces connaissances, pour ensuite pouvoir les formaliser et les organiser au sein de ressources sémantiques. Dans ce contexte, la tâche d'identification de relations est une étape cruciale, car située en amont d'autres tâches comme l'expansion sémantique en recherche d'information ou encore l'extraction de relations pour la construction de ressources sémantiques (thesaurus, taxonomies, ontologies, ressources termino-ontologiques, etc. (Buitelaar *et al.*, 2005). De nombreux travaux se sont employés à extraire les relations d'hypernonymie car elles constituent l'ossature principale de la plupart de ces types de ressources.

Deux paradigmes organisent ce champ d'étude (Tanev & Magnini, 2008; Granada, 2015) : les approches qualifiées de linguistiques font appel à des patrons pour identifier des indices de relation entre termes (Hearst, 1992) ; les approches statistiques, aujourd'hui dominantes, recourent à des procédures d'apprentissage supervisé (Pantel & Pennacchiotti, 2008) ou non-supervisé (Banko *et al.*, 2007), ou font appel à des indices distributionnels (Lenci & Benotto, 2012). Ces travaux ont appliqué ces différentes méthodes à une langue, un domaine donné (général ou de spécialité), un genre de corpus (encyclopédique, scientifique, journalistique, etc.), selon la nature des sources de connaissances utilisées (documents structurés, semi-structurés ou non structurés), ou selon l'utilisation visée de la taxonomie (intégration dans des ressources plus complexes comme des thésaurus, des termino-ontologies ou des ontologies "riches").

L'étude que nous menons, et dont nous présentons ici une première évaluation, vise à montrer l'intérêt d'appliquer différentes approches sur un même corpus pour identifier la relation d'hyperonymie, à travers ses différents modes d'expression, selon qu'elle est formulée dans une section structurée, semi-structurée ou non structurée du texte. En effet, la relation d'hyperonymie peut être exprimée par le lexique et la structure syntaxique comme dans *Le sable est une roche sédimentaire meuble*, par une inclusion lexicale comme dans *pigeon domestique* (sous-entendu *le pigeon domestique est un pigeon*), ou encore à l'aide d'éléments de ponctuation ou de mise en forme qui se substituent aux marqueurs lexicaux comme la virgule dans *Le cheval de Troie, un mythe grec* ou encore la disposition dans les structures énumératives.

Pour ce faire, nous analysons la complémentarité de deux approches, l'une linguistique basée sur des patrons lexico-syntaxiques, et l'autre statistique basée sur de l'apprentissage supervisé. Nous les avons appliquées sur un corpus constitué des pages de désambiguïsation de Wikipedia. En effet, ces pages offrent un premier cas de figure favorable : très riches en relations d'hyperonymie, elles comportent du texte rédigé (assez minoritairement), et, pour l'essentiel, du texte peu rédigé (structure syntaxique incomplète) usant de mise en forme matérielle variée comme la ponctuation, diverses polices de caractère ou la disposition.

Ce travail s'inscrit dans le cadre du projet SemPedia [1] dont l'objectif est d'enrichir la ressource sémantique DBPedia pour le français (les ressources pour la langue française faisant défaut aujourd'hui encore), en spécifiant et implémentant un ensemble de nouveaux extracteurs Wikipedia dédiés à l'extraction de la relation d'hyperonymie.

L'article est organisé de la façon suivante. La section 2 rappelle les principaux travaux connexes à notre proposition. La section 3 présente le matériel et les méthodes mis en œuvre, à savoir la description des corpus d'apprentissage et de référence, leurs pré-traitements, et les approches d'extraction retenues. Les résultats obtenus sont présentés et discutés dans la section 4. Enfin la section 5 permet de conclure et d'indiquer les perspectives envisagées.

2 Travaux connexes

En matière d'extraction de relations, le travail pionnier des méthodes linguistiques est celui de (Hearst, 1992) qui a défini un ensemble de patrons lexico-syntaxiques spécifiques à l'hyperonymie pour l'anglais. Ce travail a été repris en français pour identifier différents types de relations (Séguéla & Aussenac-Gilles, 1999), des relations d'hyperonymie entre termes (Morin & Jacquemin, 2004), des relations de méronymie (Berland & Charniak, 1999), en intégrant

1. http ://www.irit.fr/Sempedia

progressivement des techniques d'apprentissage. Dans la deuxième famille de méthodes, les travaux de Snow *et al.* (2004) et Bunescu & Mooney (2005) appliquent des techniques d'apprentissage supervisé à un ensemble d'exemples annotés à la main. On sait que le coût de l'annotation manuelle constitue la principale limite de l'apprentissage supervisé. Une manière de pallier ce problème est l'apprentissage par supervision distante qui consiste à construire l'ensemble d'exemples à l'aide d'une ressource externe (Mintz *et al.*, 2009). Cette approche requière cependant de disposer d'une ressource sémantique offrant un taux de couverture correct du corpus. D'autres manières concernent l'apprentissage semi-supervisé ou non supervisé. Brin (1998) a utilisé une sélection de patrons pour construire l'ensemble d'exemples, mettant en œuvre une méthode fondée sur l'apprentissage semi-supervisé appelée aussi bootstrapping. (Agichtein & Gravano, 2000) et (Etzioni *et al.*, 2004) ont repris cette méthode en ajoutant des traits sémantiques pour identifier des relations entre entités nommées. L'apprentissage non supervisé, basé sur des techniques de clustering, a été mis en œuvre par (Yates *et al.*, 2007) et (Fader *et al.*, 2011) qui ont utilisé des traits syntaxiques pour entraîner leurs classifieurs et identifier des relations entre entités nommées.

Au-delà de ces travaux, qui proposent et évaluent des approches une à une, peu de résultats existent sur les apports respectifs et la complémentarité de plusieurs méthodes en vue d'en optimiser la combinaison. Granada (2015) a comparé les performances de différents méthodes (par patrons, par inclusion lexicale, distributionnelles) pour la tâche d'extraction de la relation d'hyperonymie dans différentes langues, en définissant plusieurs métriques telles que, outre les mesures classiques d'évaluation (rappel et précision), la densité et la profondeur des hiérarchies. L'évaluation a été menée sur différents types de corpus mais ne prend pas en compte les approches par apprentissage.

Dans (Yap & Baldwin, 2009), les auteurs étudient l'impact du choix du corpus, la taille des exemples d'entraînement sur la performance de méthodes de même nature (méthodes supervisées), sur plusieurs types de relation (hyperonymie, synonymie et antonymie), tandis que dans (Ben Abacha & Zweigenbaum, 2011), une approche hybride permet d'extraire des relations entre entités spécifiques (maladie et traitement) dans des corpus biomédicaux. Cette approche repose sur des patrons, basés sur l'expertise humaine, et sur une méthode d'apprentissage statistique basée sur le classifieur SVM. L'approche calcule automatiquement les poids pour les deux différentes méthodes et ces poids sont ensuite appliqués pour intégrer la sortie de chaque méthode. Dans cette même ligne, nous exploitons des méthodes de nature différente, mais nous nous concentrons sur un type de relation spécifique.

En ce qui concerne l'enrichissement de la base de connaissances DBPedia, plusieurs applications, appelées "extracteurs" ont été développées pour analyser les différents éléments présents dans les pages Wikipédia. Ainsi Morsey *et al.* (2012) ont développé 19 extracteurs qui extraient des entités et des relations entre entités identifiés au sein de chaque élément de structure de ces pages : résumé, images, infobox, etc. D'autres travaux ont été proposés pour l'extraction de relations à partir de ces différents éléments, notamment pour les relations d'hyperonymie. Citons Suchanek *et al.* (2007) qui ont exploité la partie Catégorie dans les pages Wikipédia pour construire la base de connaissances Yago, Kazama & Torisawa (2007) qui ont exploité la partie Définition, et enfin Sumida & Torisawa (2008) qui se sont intéressés aux menus. On constate donc que la base de connaissances DBPedia est construite uniquement à partir des éléments de structure des pages Wikipédia : les travaux visant l'extraction des relations à partir de textes

n'ont pas été mis à profit pour l'alimentation de ces ressources, ce qui veut dire que la majorité des connaissances présentes dans ces pages restent inexploitées.

Nous proposons ici d'analyser la complémentarité de deux méthodes d'extraction de relations d'hypéronymie, de nature différente, afin de mieux exploiter les différentes sections du texte présentant des types de rédaction et des niveaux de structuration variables. Ces méthodes ont été appliquées sur un corpus de pages de désambiguïsation Wikipédia. Ces travaux sont décrits dans la section suivante.

3 Données et Méthodes

Cette section est consacrée à la description des corpus utilisés, des pré-traitements effectués sur ces corpus, et les méthodes d'extraction de relations que nous avons retenues.

3.1 Corpus

Des pages de nature différente peuvent être identifiés au sein de l'encyclopédie Wikipedia. Parmi elles, les pages d'homonymie (appelées aussi *pages de désambiguïsation*) listent les articles dont le titre est polysémique, et donnent une définition de toutes les acceptions recensées pour ce titre, qui renvoient à autant d'entités. Grâce aux consignes de rédaction de Wikipedia, qui recommandent l'usage de templates (*Toponymes*, *Patronymes*, etc.), ces pages présentent des régularités aussi bien rédactionnelles que de mise en forme pour présenter les différentes acceptions du terme (comportant un ou plusieurs mots) faisant l'objet de la page. De fait, pour chaque acception, une définition de ce sens et un lien vers la page correspondante sont fournis. Or les définitions sont des objets textuels au sein desquels la relation d'hyperonymie est souvent présente (Malaisé *et al.*, 2004) (Rebeyrolle & Tanguy, 2000).

Enfin, sur ces pages, les définitions prennent des formes variées mais prévisibles. Elles comportent pour l'essentiel du texte peu rédigé (structure syntaxique incomplète), mais aussi une mise en forme matérielle riche, utilisant la ponctuation, diverses polices de caractère ou la disposition.

Par exemple, la page d'homonymie *Mercure* cite plusieurs articles, et donne pour chacun une définition (Figure 1). Chaque définition présente plusieurs relations d'hyperonymie, exprimées par le lexique (*le mercure est un élément chimique*), à l'aide de caractères de ponctuation (la virgule dans *le Mercure, un fleuve du sud de l'Italie*), ou de la disposition comme dans les structures énumératives verticales (*la diode à vapeur de mercure est un appareil de mesure*, *la pile au mercure est un appareil de mesure*, etc.)

Ces pages d'homonymie étant pertinentes pour notre étude, nous avons constitué un corpus regroupant toutes celles de Wikipedia en français (dump 2016 de la version XML), soit 5924 pages de désambiguïsation. De ce corpus ont été extraits deux sous-corpus.

— 20 pages de désambiguïsation choisies aléatoirement forment le *corpus de référence*. Dans ces pages, les relations d'hyperonymie ont été annotées manuellement, en marquant les mots renvoyant aux entités en relation et la zone de texte signifiant la relation. Ce sous-corpus sera utilisé pour évaluer l'enrichissement potentiel de la ressource DBPedia ;

— les pages restantes (5904) forment un sous-corpus que nous appelons *corpus d'apprentissage*, destiné à entraîner et à évaluer notre modèle d'apprentissage (section 3.3.2).

Physique et chimie [modifier | modifier le code]

- Le mercure (symbole Hg) est un élément chimique.
- Le terme mercure rouge désignait au XIXe siècle l'iodure de mercure. Dans la dernière partie du XXe siècle il a été appliqué à une substance imaginaire, présentée comme un matériau stratégique rentrant dans la construction des armes nucléaires.
- Le millimètre de mercure (symbole mmHg), ou torr, est unité de mesure de pression.
- Plusieurs appareils de mesure ou méthodes physiques font référence au mercure, dont notamment :
  - la diode à vapeur de mercure,
  - la pile au mercure,
  - la pompe à mercure,
  - le porosimètre à mercure (en),
  - le thermomètre à mercure.

Toponyme et hydronyme [modifier | modifier le code]

Mercure est un nom de lieu notamment porté par :

- Mercure, une station du métro de Lille Métropole ;
- le Mercure, un fleuve du sud de l'Italie ;
- les îles Mercure, un archipel néo-zélandais, au large de la péninsule de Coromandel.
- le lac Mercure, un lac de l'île principale de l'archipel des Kerguelen, dans les Terres australes et antarctiques françaises ;
- le monastère Saint-Mercure, un important monastère féminin copte orthodoxe, situé dans le vieux Caire (Égypte) ;
- le mont Mercure, une montagne d'Italie ;
- Saint-Michel-Mont-Mercure, une ancienne commune française située dans le département de la Vendée, en région Pays-de-la-Loire
- la Vallée du Mercure, un grand bassin fluvial italien situé dans le sud de la Basilicate et le nord de la Calabre, et qui fut occupé par u lac au Pliocène.

FIGURE 1 – Extrait de la page de désambiguïsation *Mercure*

3.2 Pré-traitements

Chaque corpus est étiqueté morpho-syntaxiquement par TreeTagger [2]. Pour identifier l'expression de relations sémantiques, le texte est aussi annoté à l'aide de termes, à savoir des syntagmes, le plus souvent nominaux, pouvant désigner des entités ou des classes conceptuelles. Par exemple, *Mercure, système solaire, planète* sont quelques-uns des termes présents sur la figure 1. Les termes peuvent donc être inclus les uns dans les autres (*système* dans *système solaire*). Plutôt que d'utiliser un extracteur de termes, nous avons choisi de construire a priori deux listes de termes :

— LBabel comporte les labels en français des concepts présents dans la ressource sémantique BabelNet [3] ; elle servira à annoter le corpus d'apprentissage ;
— LCorpus, composée des termes annotés manuellement dans le corpus de référence ; elle sera utilisée pour l'évaluation de l'approche sur le corpus de référence.

L'annotation du corpus d'apprentissage par des termes provenant d'une source sémantique partagée assure une plus grande validité du modèle d'apprentissage. Cela évite aussi que l'identification des termes vienne biaiser l'extraction des relations.

2. http ://www.cis.uni-muenchen.de/schmid/tools/TreeTagger
3. http ://babelnet.org/

3.3 Approches retenues

Nous avons retenu deux approches, souvent opposées par le coût de leur mise en œuvre et par les résultats de précision et de rappel qu'elles fournissent : l'utilisation de patrons lexico-syntaxiques, et une approche statistique basée sur de l'apprentissage supervisé mettant en œuvre le principe de supervision distante. Alors que les patrons représentent des schémas langagiers récurrents faisant appel au lexique, à la syntaxe et à la ponctuation, l'apprentissage automatique permet de combiner de nombreux indices du corpus, de différentes natures (morphologiques, syntaxiques, sémantiques ou encore de mise en forme, ...) et de capter ainsi de façon plus globale les propriétés des contextes.

3.3.1 Patrons lexico-syntaxiques

Un patron lexico-syntaxique décrit une expression régulière, formée de mots, de catégories grammaticales ou sémantiques, et de symboles, visant à identifier des fragments de texte conformes à cette expression. Dans le cas de la recherche de relations, le patron caractérise un ensemble de formes linguistiques dont l'interprétation est relativement stable et correspond à une relation sémantique entre termes (Rebeyrolle & Tanguy, 2000). Un patron est d'autant plus efficace qu'il est adapté au corpus. Toutefois, la mise au point étant coûteuse, il est classique d'implémenter des patrons génériques comme ceux de (Hearst, 1992).

Nous avons mené deux expériences séparées, en utilisant tout d'abord **PatronsG**, une liste de 30 patrons génériques issus des travaux de (Jacques & Aussenac-Gilles, 2006), puis **PatronsGS**, cette même liste augmentée de patrons spécifiques à notre corpus et qui ont été définis manuellement [4] (Ghamnia, 2016).

3.3.2 Approche par apprentissage supervisé

Nous avons choisi de reprendre le principe de la supervision distante proposée par (Mintz *et al.*, 2009). Comme pour tout apprentissage supervisé, il convient de créer un ensemble d'exemples, d'entraîner un modèle statistique sur ces exemples, et d'évaluer le modèle sur un ensemble de test ou par validation croisée. L'originalité de la supervision distante réside dans le fait de construire les exemples automatiquement à partir d'une ressource externe. Dans le cadre de l'extraction de relations, la construction d'un exemple consiste à extraire une paire de termes d'une unité textuelle, à associer un ensemble de valeurs de traits préalablement définis, et à associer une classe. La classe correspond à la relation (si elle existe) qui lie, dans la ressource externe, deux concepts ayant pour labels les deux termes extraits du texte. Une fois entraîné à partir des exemples classés, un algorithme de classification multi-classes permet d'associer une classe à chaque exemple d'un nouveau corpus.

Nous avons adapté cette méthode en nous focalisant sur la relation d'hypéronymie, et en procédant à une classification binaire. Pour chaque exemple, nous avons défini un ensemble de propriétés exploitées par le système d'apprentissage : i) une paire de termes (ci-après Terme1 et Terme2) extraits d'une même phrase, ii) un contexte (ou fenêtre) de taille n formé par la séquence {n tokens précédant Terme1, Terme1, tokens séparant Terme1 et Terme2, Terme2, n

4. Une implémentation en JAPE de ces patrons est visible sur le site : https ://github.com/aghamnia/SemPediaPatterns

tokens suivant Terme2}, iii) un ensemble de valeurs de traits ayant des niveaux de granularité différents (voir Table 1), et iv) la classe (positif ou négatif) à laquelle appartient l'exemple. Un exemple est positif (resp. négatif) si les deux termes dénotent deux concepts qui existent dans la ressource sémantique, et si la relation d'hyperonymie entre ces deux concepts est représentée (resp. n'est pas présente) dans cette ressource. Dans tous les autres cas, la paire de termes ne constitue pas un exemple d'apprentissage.

Niveau de granularité	Trait	Signification	Type
Token	POS	Part Of Speech	chaîne de caractères
	LEMME	Forme lemmatisée du token	chaîne de caractères
Fenêtre	distT1	Nombre de tokens entre le mot et Terme1	entier
	distT2	Nombre de mots entre le token et Terme2	entier
	distT1T2	Nombre de tokens entre Terme1 et Terme2	entier
	nbMotsFenêtre	Nombre de tokens dans la fenêtre	entier
Phrase	nbMotsPhrase	Nombre de tokens dans la phrase	entier
	presVerbe	Présence d'une forme verbale	booléen

TABLE 1 – Ensemble des traits associés à un exemple.

Les mauvaises performances fournies par les analyseurs syntaxiques lorsqu'ils sont appliqués sur du texte peu rédigé nous a conduit à ne pas prendre en compte les dépendances syntaxiques.

Nous décrivons ci-dessous la construction d'un exemple à partir de la phrase *"Lime ou citron vert, le fruit des limettiers : Citrus aurantiifolia et Citrus latifolia"*

La projection de la liste des termes LBabel conduit à annoter la phrase par les termes `Lime`, `citron`, `citron vert`, `vert`, `fruit`. Considérons le couple <`Lime`, `fruit`> choisi aléatoirement par le système : Terme1=`Lime` et Terme2=`fruit`. Pour une fenêtre égale à 3 tokens [5], le système extrait alors de la phrase :

`Terme1 ou citron vert, le Terme2 des limettiers :`

où les mots correspondant aux termes ont été remplacés par Terme1 et Terme2. L'annotation par Tree-Tagger permet de remplacer les formes exactes des tokens par leur lemme précédé de leur catégorie syntaxique :

```
Terme1 KON/ou NOM/citron ADJ/vert PUN/, DET:ART/le Terme2
PRP:det/du NOM/limettier PUN/:
```

Enfin, des fonctions de traits associent à chaque token les distances relatives (en nombre de tokens) de ce token à Terme1 et à Terme2 sous forme de couple de valeurs, le nombre de tokens entre Terme1 et Terme2 (en l'occurrence 5) et le nombre de tokens dans la phrase (ici 16). Le dernier trait indique l'absence de verbe dans la phrase.

```
(0,6) (-1,5) (-2,4) (-3,3) (-4,2) (-5,1) (-6,0) (-7,-1) (-8,-2)
(-9,-3) 5 16 false
```

5. Nous avons évalué des fenêtres de dimension 1, 3 et 5, l'optimum étant obtenu pour la dimension 3

Voici l'exemple dans son intégralité :

```
Terme1 KON/ou NOM/citron ADJ/vert PUN/, DET:ART/le Terme2
PRP:det/du NOM/limettier PUN/:
(0,6) (-1,5) (-2,4) (-3,3) (-4,2) (-5,1) (-6,0) (-7,-1) (-8,-2)
(-9,-3) 5 16 false
```

Cet exemple est positif car les termes "lime" et "fruit" renvoient à des ressources en relation d'hyperonymie dans BabelNet.

Nous avons ainsi produit automatiquement ~8000 exemples, et avons conservé 6000 exemples (3000 positifs et 3000 négatifs). L'ensemble d'entraînement est composé de 4000 exemples pris aléatoirement parmi les 6000, en maintenant une quasi-parité positifs / négatifs (~2000/~2000), et l'ensemble de test comporte les 2000 exemples restants. Nous avons entraîné un algorithme de régression logistique binaire, MaxEnt (Berger *et al.*, 1996) sur l'ensemble d'entraînement. Appliqué à l'ensemble de test, MaxEnt a fourni un rappel de 0.63 et une précision de 0.71.

4 Expériences, résultats et discussion

A partir du corpus de référence, 688 exemples vrais positifs (VP) et 267 exemples vrais négatifs (VN) ont été identifiés manuellement. Nous interprétons les résultats obtenus afin de juger de la complémentarité des méthodes expérimentées, et de l'intérêt de leur utilisation conjointe.

4.1 Évaluation quantitative

Nous avons mis en œuvre les deux approches indépendamment, l'approche par patrons avec **PatronsG** puis avec **PatronsGS** (voir section 3.3.1), et l'approche par apprentissage supervisé (MaxEnt). Nous avons évalué leur complémentarité à l'aide de l'union et de l'intersection de leurs résultats. Le tableau 2 fournit les différentes valeurs de la précision, du rappel, de la F-mesure et de l'exactitude.

	PatronsG	PatronGS	MaxEnt	PatronsG union MaxEnt	PatronsGS union MaxEnt
Précision	0.96	0.81	0.71	0.72	0.73
Rappel	0.04	0.46	0.63	0.65	0.77
F-mesure	0.07	0.59	0.67	0.68	0.75
Exactitude	0.31	0.54	0.55	0.56	0.63

TABLE 2 – Évaluation des approches.

Confirmant l'état de l'art (Hearst, 1992) (Malaisé *et al.*, 2004), les patrons obtiennent un fort taux de précision, au détriment d'un faible rappel. Et comme attendu, les patrons génériques augmentés des patrons spécifiques donnent de meilleurs résultats que les patrons génériques seuls : bien que la précision baisse à 0.81, le rappel passe à 0.46, pour une F-mesure de 0.59. En revanche, l'approche par apprentissage supervisé, moins bonne en précision, produit un meilleur taux de rappel. En effet, le corpus présente de fortes régularités (aussi bien syntaxiques

que de mise en forme), ce qui permet de renforcer l'apprentissage. Ces résultats corroborent les résultats reportés dans la littérature pour d'autres types de corpus.

Nous constatons que l'application des deux approches fournit une meilleure F-mesure que les approches prises indépendamment, et que l'union des résultats de PatronsGS et de MaxEnt permet d'obtenir la meilleure F-mesure. Nous avons également pu analyser la complémentarité des PatronsGS et de MaxEnt à travers les résultats donnés dans la Table 3.

	PatronsGS ou MaxEnt	PatronsGS et MaxEnt	PatronsGS seul	MaxEnt seul	Aucune des 2 méthodes
Nombre VP	527	221	96	210	161

TABLE 3 – Parmi les 688 VP du corpus de référence, nombre de VP trouvés par les approches.

Nous avons pu ainsi constater que parmi les 306 VP qui ne sont identifiés que par une seule des deux méthodes, PatronsGS en identifie 32%, et MaxEnt 68%. Ce résultat confirme la complémentarité de ces deux approches.

4.2 Evaluation qualitative

Nous avons constaté que parmi les 221 relations trouvées par PatronsGS et MaxEnt, 9 relations concernent des relations exprimées par le verbe *être*, comme entre les termes *macédoine* et *salade de fruits* dans la phrase "La macédoine est une salade de fruits ou de légumes". Quasiment toutes les autres relations correspondent au schéma "X, Y" comme dans "Le cheval de Troie, un mythe grec". On remarque toutefois que les patrons retrouvent des relations entre des noms communs, alors que MaxEnt trouve essentiellement des relations entre entités.

Pour les 96 relations trouvées par PatronsGS et n'ayant pas été identifiées par MaxEnt, on trouve des relations (19) exprimées par le verbe *être*, mais lorsque la relation n'est pas exprimée en début de phrase, comme la relation entre *Babel fish* et *espèce imaginaire* dans "Le poisson Babel ou Babel fish est une espèce imaginaire". Quasiment toutes les autres relations trouvées seulement par les patrons correspondent au schéma "X,Y" comme décrit ci-dessus. Nous n'avons pas encore identifié la cause du silence de MaxEnt à ce niveau.

Parmi les 210 relations trouvées par MaxEnt et non identifiées par PatronsGS, on trouve les cas d'inclusion lexicale (85), comme la relation entre *gare de Paris Bastille* et *gare*, issue du groupe nominal "gare de Paris Bastille". MaxEnt permet également de trouver des relations exprimées par d'autres verbes d'état (8) comme la relation entre *aigle* et *oiseaux rapaces* dans "Aigle désigne en français certains grands oiseaux rapaces". MaxEnt identifie aussi des relations exprimées dans des unités textuelles comportant des coordinations, comme la relation entre *poisson Babel* et *espèce imaginaire* dans "Le poisson Babel ou Babel fish est une espèce imaginaire", ou encore entre *Louis Babel* et *explorateur* dans "Louis Babel, prêtre-missionnaire oblat et explorateur". Enfin MaxEnt identifie très bien les relations au sein de texte usant de mise en forme comme la relation entre *arête* et *barbe de l'épi* dans "Arête, "barbe de l'épi"", ou entre *Aigle* et *chasseur de mines* dans "Aigle (M647), chasseur de mines".

Finalement, parmi les 161 relations qui n'ont été trouvées ni par PatronsGS ni par MaxEnt, 55 correspondent à des inclusions lexicales (nous n'avons pas non plus identifié la cause du silence de MaxEnt), 64 possèdent une incise entre les deux termes comme dans "Un Appelant

(jansénisme) est, au XVIIIe siècle, un ecclésiastique qui ... " ou "Giuseppe Cesare Abba (1838-1910), écrivain". Les 42 cas restants concernent des formes d'expression non prises en charge par les patrons et trop peu fréquentes pour être apprises par MaxEnt, comme "X tel que Y".

Cette analyse confirme l'intérêt de mettre en œuvre sur un même corpus des approches complémentaires. Tout d'abord, nous avons pu constater que les inclusions lexicales sont identifiées par MaxEnt seul. Ensuite, les différentes occurrences de relations au sein d'une même phrase sont identifiées par les deux méthodes, comme on l'a vu ci-dessus à travers l'exemple "Le poisson Babel ou Babel fish est une espèce imaginaire". Enfin, MaxEnt permet d'identifier les relations exprimées selon différentes variantes d'un même schéma (incises, usage de mise en forme, etc.), dès lors que ces structures sont récurrentes. Par ailleurs, nous avons également pu observer que PatronsGS et MaxEnt sont complémentaires dans une proportion de ~1/3 vs. 2/3.

4.3 Enrichissement de la ressource DBPedia

Nous avons évalué l'enrichissement de la ressource DBPedia par les relations extraites par PatronsGS et/ou par MaxEnt. Pour cela, nous avons manuellement vérifié la présence dans DBPedia des 688 relations VP, en interrogeant la ressource DBPedia pour vérifier si des entités aux labels proches de *Terme1* et *Terme2* y sont liées par un chemin formé de relations *rdf :type* et *rdf :subclassOf*. Nous avons empiriquement fixé à 3 la longueur maximale de ce chemin. Cette vérification manuelle se justifie par le fait que les termes de LCorpus peuvent différer des labels de DBPedia.

Parmi ces 688 relations, 199 relations ne sont pas exprimées dans DBPedia. 103 de ces 199 relations ont été identifiées par MaxEnt et 42 d'entre elles ont été trouvées par PatronsGS. En considérant l'union des résultats des deux approches, 125 relations identifiées (20 relations étant présentes dans l'intersection des résultats individuels) ne sont pas dans DBPedia . La Table 4 présente le taux d'enrichissement de la ressource par rapport aux relations identifiées par chaque méthode. Ces résultats confirment que les textes de Wikipedia contiennent des relations d'hyperonymie non encore exploitées par les extracteurs Wikipédia.

	PatronsGS	MaxEnt	Union PatronsGS et MaxEnt
Taux d'enrichissement	0.21%	0.51%	0.63%

TABLE 4 – Taux d'enrichissement de DBPedia.

5 Conclusion et perspectives

Ces premières expériences nous ont permis de mettre en place des données et un protocole pour la comparaison de deux méthodes d'extraction des relations, de manière à en analyser finement la complémentarité. Les premiers résultats sont encourageants et convergent avec les travaux de (Malaisé *et al.*, 2004) (Granada, 2015) (Buitelaar *et al.*, 2005). Nous envisageons de les pousser plus loin dans plusieurs directions. Nous souhaitons en priorité intégrer d'autres techniques pour prendre en compte d'autres éléments textuels, par exemple le système qui traite les structures énumératives verticales et régulières (Kamel & Trojahn, 2016) ou encore les outils développés dans (Granada, 2015). Pour améliorer les performances de chaque méthode, outre un meilleur encodage des patrons, nous prévoyons de rajouter des nouveaux traits pour

l'apprentissage automatique. Bien sûr, la méthode devra être testée sur un autre corpus incluant d'autres types de pages Wikipedia.

A terme, notre ambition est de croiser les méthodes de manière à ce que les résultats des unes servent d'entrées plus riches aux autres, et améliorent ainsi leurs performances. La première piste envisagée dans ce sens serait d'utiliser les patrons pour annoter le corpus permettant d'indiquer qu'un patron a été (ou non) reconnu dans le contexte de deux termes, ce qui serait un signe fort de présence de la relation marquée par ce patron. Ce type de trait permettrait d'entraîner le classifieur à reconnaître plusieurs types de relations en plus de l'hyperonymie.

Références

AGICHTEIN E. & GRAVANO L. (2000). Snowball : Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, p. 85–94 : ACM.

BANKO M., CAFARELLA M. J., SODERLAND S., BROADHEAD M. & ETZIONI O. (2007). Open information extraction from the web. In *IJCAI*, volume 7, p. 2670–2676.

BEN ABACHA A. & ZWEIGENBAUM P. (2011). *A Hybrid Approach for the Extraction of Semantic Relations from MEDLINE Abstracts*, In A. GELBUKH, Ed., *Computational Linguistics and Intelligent Text Processing : 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part II*, p. 139–150. Springer Berlin Heidelberg.

BERGER A. L., PIETRA V. J. D. & PIETRA S. A. D. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, **22**(1), 39–71.

BERLAND M. & CHARNIAK E. (1999). Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, p. 57–64 : Association for Computational Linguistics.

BRIN S. (1998). Extracting patterns and relations from the world wide web. In *International Workshop on The World Wide Web and Databases*, p. 172–183 : Springer.

BUITELAAR P., CIMIANO P. & MAGNINI B. (2005). Ontology learning from text : An overview. In *Ontology Learning from Text : Methods, Evaluation and Applications*, p. 3–12 : IOS Press.

BUNESCU R. C. & MOONEY R. J. (2005). A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, p. 724–731 : Association for Computational Linguistics.

ETZIONI O., CAFARELLA M., DOWNEY D., KOK S., POPESCU A.-M., SHAKED T., SODERLAND S., WELD D. S. & YATES A. (2004). Web-scale information extraction in knowitall :(preliminary results). In *Proceedings of the 13th international conference on World Wide Web*, p. 100–110 : ACM.

FADER A., SODERLAND S. & ETZIONI O. (2011). Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 1535–1545 : Association for Computational Linguistics.

GHAMNIA A. (2016). Extraction de relations d'hyperonymie à partir de wikipédia. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016*.

GRANADA R. L. (2015). *Evaluation of methods for taxonomic relation extraction from text*. PhD thesis, Pontifícia Universidade Católica do Rio Grande do Sul.

HEARST M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, p. 539–545, Stroudsburg, PA, USA : Association for Computational Linguistics.

JACQUES M.-P. & AUSSENAC-GILLES N. (2006). Variabilité des performances des outils de TAL et genre textuel. Cas des patrons lexico-syntaxiques. *Traitement Automatique des Langues, Non Thématique*, **47**(1), (en ligne).

KAMEL M. & TROJAHN C. (2016). Exploiter la structure discursive du texte pour valider les relations candidates d'hyperonymie issues de structures énumératives parallèles. In *IC 2016 : 27es Journées francophones d'Ingénierie des Connaissances (Proceedings of the 27th French Knowledge Engineering Conference), Montpellier, France, June 6-10, 2016.*, p. 111–122.

KAZAMA J. & TORISAWA K. (2007). Exploiting wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 698–707.

LENCI A. & BENOTTO G. (2012). Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation*, p. 75–79 : Association for Computational Linguistics.

MALAISÉ V., ZWEIGENBAUM P. & BACHIMONT B. (2004). Detecting semantic relations between terms in definitions. In S. ANANADIOU & P. ZWEIGENBAUM, Eds., *COLING 2004 CompuTerm 2004 : 3rd International Workshop on Computational Terminology*, p. 55–62, Geneva, Switzerland : COLING.

MINTZ M., BILLS S., SNOW R. & JURAFSKY D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 2-Volume 2*, p. 1003–1011 : Association for Computational Linguistics.

MORIN E. & JACQUEMIN C. (2004). Automatic acquisition and expansion of hypernym links. *Computers and the Humanities*, **38**(4), 363–396.

MORSEY M., LEHMANN J., AUER S., STADLER C. & HELLMANN S. (2012). Dbpedia and the live extraction of structured data from wikipedia. *Program*, **46**(2), 157–181.

PANTEL P. & PENNACCHIOTTI M. (2008). Automatically harvesting and ontologizing semantic relations. *Ontology learning and population : Bridging the gap between text and knowledge*, p. 171–198.

REBEYROLLE J. & TANGUY L. (2000). Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de Grammaire*, **25**, 153–174.

SÉGUÉLA P. & AUSSENAC-GILLES N. (1999). Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. In *Conférence ingénierie des connaissances*, p. 79–88.

SNOW R., JURAFSKY D. & NG A. Y. (2004). Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17*.

SUCHANEK F. M., KASNECI G. & WEIKUM G. (2007). Yago : A core of semantic knowledge unifying wordnet and wikipedia. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, p. 697–706.

SUMIDA A. & TORISAWA K. (2008). Hacking wikipedia for hyponymy relation acquisition. In *IJCNLP*, volume 8, p. 883–888 : Citeseer.

TANEV H. & MAGNINI B. (2008). Weakly supervised approaches for ontology population. In *Proceeding of the 2008 conference on Ontology Learning and Population : Bridging the Gap between Text and Knowledge*, p. 129–143 : Citeseer.

YAP W. & BALDWIN T. (2009). Experiments on pattern-based relation learning. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, p. 1657–1660 : ACM.

YATES A., CAFARELLA M., BANKO M., ETZIONI O., BROADHEAD M. & SODERLAND S. (2007). Textrunner : open information extraction on the web. In *Proceedings of Human Language Technologies : The Annual Conference of the North American Chapter of the Association for Computational Linguistics : Demonstrations*, p. 25–26 : Association for Computational Linguistics.

Ontologie et TALN : l'anonymisation au service du repérage conceptuel dans le contexte de la SLA

Sonia Cardoso[1], Luis Felipe Melo Mora[2], Marie-Christine Jaulent[2], Xavier Aimé, David Grabli[3], Vincent Meininger[5], Jean Charlet[2,4]

[1] IHU-A-ICM Institut des Neurosciences Translationnelles de Paris,
s.cardoso-ihu@icm-institute.org
[2] INSERM UMRS 1142, LIMICS, F-75006, Paris
Sorbonne Universités, UPMC Univ. Paris 06, UMR_S 1142, LIMICS, F-75006, Paris
Université Paris 13, Sorbonne Paris Cité, LIMICS, (UMR_S 1142), F-93430, Villetaneuse
luisfe.melo@gmail.com
[3] Assistance Publique Hôpital Pitié Salpêtrière, Département des maladies du Système Nerveux, Paris
Université Pierre et Marie Curie
david.grabli@psl.aphp.fr
[4]Assistance Publique –Hôpitaux de Paris DRCD, F-75004 PARIS
jean.charlet@upmc.fr
[5]Ramsay General de Santé, Hôpital Peupliers Paris
vincent.meininger@psl.aphp.fr

Résumé : L'objectif de notre travail est l'exploitation de la base évènementielle du réseau SLA (Sclérose Latérale Amyotrophique) d'Île-de-France (IDF), pour permettre à long terme, de comprendre les ruptures de parcours de santé. Pour analyser ce corpus une chaine de pré traitement est nécessaire. L'un de ces processus est l'anonymisation, processus consistant à masquer l'ensemble des éléments ne permettant pas l'identification d'une personne. Ce processus de changement de données nominales en catégories sémantiques, permet secondairement une amélioration du repérage des concepts de l'ontologie du domaine, lors de l'utilisation d'outils du traitement automatique de la langue naturelle (TALN).

Mots-clés : Ontologie, anonymisation, parcours de soins, sclérose latérale amyotrophique.

1 Introduction

L'Ingénierie des Connaissances permet la construction d'ontologies notamment dans le domaine médical qui, associées aux outils de Traitement Automatique de la Langue Naturelle (TALN), permettent d'exploiter des corpus à des fins de compréhension de processus et d'analyse.

L'objectif de notre travail, à long terme, est d'analyser et identifier les indicateurs de ruptures dans le parcours de santé[1] de personnes ayant une pathologie neurodégénérative, en particulier la Sclérose Latérale Amyotrophique (SLA) en exploitant la base de données « évènementielle »

[1] L'Article 14 de la Loi de modernisation de notre système de santé définit dans l'article L. 6327-1 du Code de la santé publique, le parcours de santé complexe, lorsque l'état de santé, le handicap ou la situation sociale du patient rend nécessaire l'intervention de plusieurs catégories de professionnels de santé, sociaux ou médico-sociaux

crée dans le cadre du réseau SLA IDF[2]. Ce réseau est représentatif du suivi de ces patients, puisqu'il accompagne 92% des patients SLA en Île de France (Cordesse *et al.*,2015).

A ce jour la base contient 2245 dossiers patients soit plus de 35 000 évènements. Les événements de coordination sont sous forme de données textuelles non structurées organisées chronologiquement. Ces derniers sont polymorphes dans leurs structures et contenus, ils peuvent contenir des demandes émises par les agents (patients, familles, professionnels), les réponses et actions mises en place par les coordinateurs et des informations de type comptes rendus médicaux.

Pour atteindre l'objectif visé, différentes étapes préliminaires sont nécessaires au traitement des corpus. Nous présenterons les étapes réalisées jusqu'à maintenant. Dans la section 2, nous décrivons le processus de construction d'une ontologie du domaine avec les diverses réflexions menées lors de cette étape pour *(i)* le choix des concepts en lien avec les ontologies et classifications existantes, *(ii)* le choix des labels préférés et synonymes spécifiques aux corpus. La section 3 concernera le processus d'anonymisation et les outils utilisés (plus spécifiquement leur apport dans le repérage d'entités. La section 4 nous permettra de replacer ce travail dans le contexte de l'architecture de traitement mise en œuvre et nous terminerons (section 5) par les limites de ce travail et les perspectives que nous envisageons.

2 Méthodologie de construction de l'ontologie

De nombreuses ontologies ont été réalisées dans le domaine de la médecine (comme par exemple *Ontolurgences* (Charlet *et al.*, 2012), *Bilingual Ontology of Alhzeimer's disease and Related Diseases*, ONTOAD (Dramé *et al.*, 2014), ou bien dans le domaine de la coordination des soins infirmiers comme la *Nursing Care Coordination Ontology*[3], NCCO (Popejoy *et al.*, 2014)). Cependant aucune à ce jour ne regroupe les maladies neurodégénératives, la coordination de parcours de santé ou les aspects sociaux et médico-sociaux spécifique au système français.

Nous nous sommes inspirés de la méthode ARCHONTE développée par B. Bachimont (2002) pour construire notre ontologie.

Une première modélisation des actes de coordination fut réalisée en utilisant les diagrammes de séquence ayant « pour but de décrire les modalités de communication entre les objets d'une application, d'un processus ou d'une organisation » (UML 2, 2006). La coordination de parcours de santé consiste en l'interaction d'*agents* (patients, neurologues coordinateurs SLA, …) réalisant des *actions* (demandes, communications…) à destination d'autres *agents* afin d'intervenir sur des *états* ou des *objets* dans des lieux spécifiés (Cardoso *et al.*, 2016).

Le choix et l'agencement des concepts s'est fait en tenant compte de modèles et de classifications existants comme l'ICF (International Classification of Functioning, Disability and Health)[4]. Ces concepts concernent les activités de vie quotidienne, les fonctions de l'organisme ou encore la Classification et terminologie des produits d'assistance pour personnes en situation de handicap (ISO 9999) dans le cadre des aides techniques.

Pour chaque concept, un travail sur deux axes fut mené. Tout d'abord (1) une recherche dans l'ensemble des corpus des synonymes, acronymes, abréviations désignant un même concept. En effet, le coordinateur peut utiliser différentes formes d'un même terme. Par exemple le concept de *Médecin Traitant* est transcrit de huit manières différentes : *méd traitant, MT, MDT, med tt, méd traitant, med ttt* ou encore *médecin de famille*. Afin d'harmoniser la saisie des abréviations et des acronymes utilisés, un travail collaboratif fut mené avec les coordinateurs aboutissant à la création de listes définies. Ces listes sont utilisées secondairement dans le travail d'anonymisation.

[2] http://reseau-sla-idf.fr

[3] http://purl.bioontology.org/ontology/NCCO

[4] http://apps.who.int/classifications/icfbrowser/

Le second axe de travail (2) est l'alignement d'une partie des concepts (travail toujours en cours) avec les classifications et ontologies françaises existantes comme ONTOPYSCHIA (Richard *et al.*, 2013), ONTOLURGENCES (Charlet *et al.*, 2012) et MENELAS (Charlet *et al.*, 2012. Nous avons également utilisé la plateforme HeTOP (Health Terminology / Ontology Portal)[5] qui contient plus de 222 800 définitions (Grosjean *et al.*, 2011) et regroupe un ensemble de terminologies et ontologies du domaine médical et les identifiants Unified Medical Language System (UMLS).

Il nous a semblé nécessaire d'utiliser des classifications et ontologies de référence car notre objectif à long terme est d'appliquer les outils que nous développons à d'autres bases événementielles utilisées pour la coordination notamment la maladie de Parkinson. A ce jour, l'ontologie est constituée de 2946 concepts et nous travaillons à la désignation et à la mise en place des relations entre les concepts.

L'ontologie créée est utilisée secondairement dans le système GATE pour servir de ressource linguistique, permettant le repérage et l'annotation des concepts dans les corpus.

3 Travail d'anonymisation

Pour exploiter les corpus il est nécessaire d'anonymiser les données comme le recommande la Commission Nationale Informatique & Libertés (CNIL) et les autorités de protection des données européennes[6] (2014). Ce processus implique de retirer suffisamment d'éléments pour que la personne concernée ne puisse plus être identifiée. Les évènements rédigés sont nominatifs (nom, prénom), temporalisés (date), et localisés (ville, nom de structure). Les informations à anonymiser sont nombreuses. Pour ce travail spécifique, nous avons collaboré avec le LIMSI[7] afin d'utiliser les outils d'annotations qu'ils ont développés : le système d'apprentissage statistique WAPITI[8] (Lavergne *et al.*, 2010) et un système à base de règles et de lexiques DARK (Data Annotation using Rules and Knowledge)[9].

Nous avons défini vingt-deux catégories faisant référence au type sémantique des données à anonymiser dans les corpus regroupées en cinq catégories principales : (1) les *agents* qui concernent des personnes physiques et indiquent leurs « fonctions » *patients, entourage, neurologues* ; (2) les *dates* qui regroupent l'ensemble des éléments temporelles (jour mois années) ; (3) les *structures* pour les *associations de patients*, les *hôpitaux*, les *structures médico-sociales* ; (4) les *identifiants numériques* pour les *numéros de dossiers*, *numéro de téléphone* ; (5) les *lieux* faisant référence à une localisation spatiale : *pays*, *villes*, *départements* et *adresse*.

3.1 Outils d'anonymisation utilisés et résultats

A l'image des approches appliquées en fouille de textes, les approches utilisées en désidentification automatique reposent sur deux grandes familles de méthodes : les méthodes à base de règles (généralement implémentées sous la forme d'expressions régulières) et de listes (listes d'entités, dictionnaires, etc..), dites « méthodes symboliques » et les méthodes à base d'apprentissage statistiques reposant sur le repérage des entités à partir d'un corpus annoté (Grouin, 2013). La première étape a consisté à anonymiser manuellement les données de cinquante-quatre dossiers (soit 2311 évènements) extraits de la base de façon aléatoire. Dans un second temps, ces dossiers anonymisés ont servis de référentiel pour le système d'apprentissage WAPITI, afin de créer un modèle d'apprentissage pour annoter

[5] http://www.hetop.eu/hetop/

[6] Groupe de travail « Article 29 » sur la protection des données. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

[7] https://www.limsi.fr/fr/

[8] https://wapiti.limsi.fr

[9] https://perso.limsi.fr/grouin/inalco/1617/index.html

automatiquement de nouveaux corpus. Il est apparu rapidement des difficultés et des erreurs en utilisant le système d'apprentissage seul, du fait des spécificités des corpus non structurés et hétérogènes. En particulier, n'étaient pas détectés correctement le type des personnes (*nom de patient annoté comme professionnel*) et l'oubli de certaines dates et lieux. Pour diminuer ces oublis et erreurs il fut décidé d'utiliser, en complément du système WAPITI (système à base d'apprentissage), un outil à base de règles, DARK, utilisant des lexiques et expressions régulières. Nous avons créé 13 listes d'expressions régulières (nom de patient[10], liste de ville et de pays, …) afin de fournir au système les termes à anonymiser.

Un nouveau corpus constitué de 15 dossiers (soit 431 évènements), tirés de façon aléatoire sur la base, fut testé pour mesurer la performance des deux systèmes en utilisant successivement (1) le système par apprentissage (WAPITI) puis (2) le système à base de règles (DARK). Les résultats obtenus donnèrent un rappel de 0,81 ainsi qu'une précision de 0,90 et une F-mesure de 0,85. Les erreurs obtenues correspondent à des heures annotées comme des dates du fait de leur format.

Initialement nous avions créé, dans le corpus de test, une distinction entre les *patients* et l'*entourage* car, d'un point de vue clinique, il est important de comprendre ce qui émane du patient – ses choix de vie, ses demandes, ses besoins – et ce qui émane de l'entourage (famille, amis). Cependant les outils d'anonymisation utilisés, en particulier l'emploi de lexique (liste des noms de patients) dans le système à base de règles ne permettent pas de faire cette distinction, ces éléments devront être pris en compte lors de l'analyse clinique des parcours.

3.2 Utilisation du corpus anonymisé

Une fois le corpus anonymisé, des traitements doivent être effectués pour permettre le repérage des concepts, des informations contextuelles sur ces concepts et des relations existants entre eux. Pour cela nous utilisons des outils développés sur la plateforme GATE[11] pour annoter les corpus, intégrant des applications utilisant l'ontologie du domaine comme ressource lexicale. Les ontologies et les outils de TALN permettent ensuite un travail de recherche intéressant comme le souligne Bodenreider (2008) : « The terminological component of biomedical ontologies is an important resource for natural language processing systems [45] and supports knowledge management tasks such as annotation (or indexing) of resources, information retrieval, access to information and mapping across resources. » et comme cela est confirmé par Charlet *et al.* (2015).

Lors du travail d'anonymisation nous avons émis l'hypothèse, que ce processus aurait un impact, qualitatif et quantitatif, sur le repérage des entités nommées lors de l'utilisation de GATE, en les remplaçant directement par les catégories les subsumant. La transformation d'une donnée nominale (JEAN DUPONT), non défini dans l'ontologie, en donnée identifiant un concept l'*agent* (*Patient*) défini dans l'ontologie permet d'un point de vue conceptuel de repérer les interactions entre les *agents* et les *actions*. Le tableau 1 illustre ces modifications.

Corpus original	Appel de Jean DUPONT qui a sollicité Claire MARCHE pour obtenir un certificat médical du Dr MUSCLE pour son dossier MDPH[12], souhaite avoir des nouvelles.
Corpus anonymisé	Appel de ***Patient*** qui a sollicité ***Coordinateur SLA*** pour obtenir un certificat médical du ***Neurologue*** pour son dossier MDPH, souhaite avoir des nouvelles.

Tableau 1 *- Exemple d'un événement transformé par l'anonymisation.*

D'un point de vue quantitatif, l'anonymisation permet d'augmenter le nombre de concepts annotés. Nous passons ainsi de 2908 concepts annotés dans le dossier initial à 3578 concepts pour le dossier anonymisé.

[10] Il faut noter que cette liste est évidemment disponible dans le logiciel d'enregistrement des événements et que, dans l'optique d'une mise en œuvre de notre système en routine, elle pourra être remise à jour régulièrement.

[11] https://fr.wikipedia.org/wiki/Architecture_générale_pour_le_traitement_de_texte

[12] Maison Départementale des Personnes Handicapées

D'un point de vue qualitatif, l'anonymisation permet de mieux repérer les « *agents* » acteurs dans le parcours de santé. En effet, la coordination de parcours de santé repose sur l'interaction d'*agents* réalisant des *actions* à destination d'autres *agents* portant sur des *objets*. L'anonymisation permet de comprendre les interactions et les interdépendances des *agents* – et des *structures* – mais aussi de la prépondérance de la présence *versus* absence de certains. En effet, le type de structures sollicitées et le type d'agents impliqués dans la coordination est un facteur important de compréhension des parcours de santé (qui sont les demandeurs, qui sont les agents donnant l'alerte sur la dégradation d'une situation clinique ou familiale comme l'épuisement).

Le processus d'anonymisation va donc avoir une action sur son objectif premier qui est de désidentifier des données afin de pouvoir exploiter des corpus à des fins de recherches, mais aussi améliorer le repérage des concepts par les outils du TALN. Bien que nécessaire et importante dans la démarche d'exploitation des corpus, l'anonymisation ne reste qu'une des étapes du flux de données mis en place dans GATE que nous décrivons dans la section suivante.

4 Traitements effectués par GATE

Les textes anonymisés sont ensuite intégrés dans le flux de données de GATE défini comme suit : (1) correction orthographique en utilisant successivement *(i)* un dictionnaire de la langue française (Hunspell), *(ii)* un dictionnaire du domaine (vocabulaire d'UMLS en français) et le lexique de notre ontologie. Une fois un mot mal orthographié repéré, un ensemble de suggestions est proposé par le système en utilisant les termes dénotant les concepts de l'ontologie du domaine et le dictionnaire Hunspell. Nous appliquons, pour ce faire, la distance d'édition Damerau-Levenshtein à chaque mot mal orthographié et ses suggestions, pour déterminer le meilleur choix. Il convient de noter que la distance d'édition ne peut pas être plus grande qu'un seuil spécifié, dans ce cas, la correction ne sera pas faite et le mot restera mal orthographié. (2) reconnaissance de concepts (définis dans l'ontologie) identifiant les extraits du texte qui font référence à ces entités. Pour mener à terme cette tâche, une lexicalisation (PoS : Part of Speech, lemme, etc.) des ressources de l'ontologie (Classes, Instances, Propriétés) est nécessaire. Par la suite, un appariement est effectué entre ces lexicalisations et des fragments textuels. (3) peuplement de l'ontologie à partir des instances repérées dans les corpus.

5 Conclusion et perspectives

Si les processus d'anonymisation et de prétraitements réalisés sur les corpus apportent des bénéfices sur l'analyse et l'exploitation des corpus d'un point de vue conceptuel, de nombreuses étapes restent encore à mener. Nous envisageons de travailler prochainement sur la prise en compte des notions hypothétiques (*il est possible que le patient rentre à domicile demain*) et niées (*le patient ne rentrera pas demain à domicile*) présentent dans les corpus, toutes deux ayant un impact sur la compréhension des évènements.

Le second axe de travail sur lequel nous souhaitons avancer est la temporalité. Actuellement, nous avons mis un élément « date » afin d'anonymiser les corpus ; cependant la temporalité est un élément important à prendre en compte lorsque l'on parle de parcours de santé. Un des axes de travail sera de décider d'un mode de représentation formel du temps dans l'ontologie.

Enfin, jusqu'à présent, l'ensemble des outils furent testés pour le corpus du réseau SLA, l'un des objectifs de notre travail est de transposer ces outils à d'autres bases de coordination neurologique. Des essais ont été faits sur la base de la maladie de Parkinson et l'hypothèse de la transposabilité semble valide, même si des éléments spécifiques devront être implémentés.

Remerciements

Nous souhaitons remercier Cyril Grouin pour sa participation active, ses outils et ces conseils dans la cadre du travail d'anonymisation.

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.

BONDENREIDER O. (2008) Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support. Yearbook of medical informatics, p 67-69

CARDOSO S., AIME X., MELO MORA L-F., JAULENT M-C., GRABLI D., MEININGER V., CHARLET J. (2016) Les ontologies pour aider à comprendre les parcours de santé dans le cadre des maladies neurodégénératives. Conférence: IA & Santé 2016 - Deuxième Atelier sur l'Intelligence Artificielle et la Santé, At Montpellier

CHARLET J., DECLERCK G., DHOMBRES F., GAYET P., MIROUX P.ET VANDENBUSSCHE P.-Y. (2012) Construire une ontologie médicale pour la recherche d'information : problématiques terminologiques et de modélisation. In : Szulman S., coordinateur. Actes des 23es Journées Ingénierie des Connaissances, Paris, France, 27-29 juin, p. 33-48.

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

CHARLET J., DARMONI S -J. (2015) Knowledge Representation and Management. From Ontology to Annotation. Findings from the Yearbook 2015 Section on Knowledge Representation and Management. Yearbook of Medical Informatics, p 134-136.

CORDESSE V., SIDOROCK F., SCHIMMEL P., HOLSTEIN J., MEININGER V. (2015) Coordinated care affects hospitlization and prognosis in amyotrophic lateral sclerosis: a cohort study. BMC Health Services Research

DRAME K., DIALLO G., DELVA F., DARTIGUES J-F., MOUILLET E., SALAMON R., MOUGIN F. (2014) Reuse of termino-ontological ressources and text corpora for building a multilingual domain ontology: An application to alzheimer's disease. Journal of Biomedical Informatics 48, 171-182

GROUIN C. (2013) Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique, PhD Thèse. Université Pierre et Marie Curie, Paris, France.

GROSJEAN, J; MERABTI, T; DAHAMNA, B; KERGOURLAY, I; THIRION B; SOUALMIA LF & DARMONI, SJ. (2011) Health Multi-Terminology Portal: a semantics added-value for patient safety. Patient Safety Informatics - Adverse Drug Events, Human Factors and IT Tools for Patient Medication Safety, Studies in Health Technology and Informatics, Volume 166, Pages 129-138.

GOMEZ-PEREZ, A., FERNANDEZ M. AND DE VICENTE A.J. (1996) "Towards a Method to Conceptualize Domain Ontologies", ECAI-96 Workshop on Ontological Engineering, Budapest.

LAVERGNE T., CAPPE O., FRANCOIS Y. (2010) Pratical VeryLarge Scale crfs. Proceedings the 48th Annual Meeting of the Association for Computational Linguistics, 504-513. Uppsal, Sweden

Popoejoy LL., Khalilia M-A., Popescu M., Galambos C., Lyons V., Rantz M., Hicks L., Stetzer F., (2014) Quantifying care coordination using natural language processing and domain-specific ontology. Journal of the American Medical Informatics Association, Volume 22, p 93-103.

PILONE D., PITMAN N. (2006) UML 2 en concentré. Edition O'Reilly, Paris.

RICHARD M., AIME X., KREBS M.-O. & CHARLET J. (2013) Au-delà du DSM : les ontologies comme aide aux classifications descriptives psychiatriques ? 2e édition du Symposium sur l'Ingénierie de l'Information Médicale, Jul 2013, Lille, France.

Peuplement d'une base de connaissance par annotation automatique de textes relatifs à la cosmétique

Molka Tounsi Dhouib[1], Cédric Lopez[2], Catherine Faron Zucker[1], Elena Cabrio[1], Fabien Gandon[1], Frédérique Segond[2]

[1] UNIVERSITÉ CÔTE D'AZUR, INRIA, CNRS, I3S, SOPHIA ANTIPOLIS, FRANCE
{dhouib,faron,cabrio}@i3s.unice.fr, fabien.gandon@inria.fr

[2] VISEO, R&D, GRENOBLE, FRANCE
{cedric.lopez,frederique.segond}@viseo.fr

Résumé : Dans cet article, nous proposons une approche pour construire une base de connaissances à partir de textes dans le domaine de la cosmétique. Il s'agit d'un cas particulier pour un domaine fixé du problème de l'extraction de relations à partir de textes. Dans le but de résoudre ce problème, nous proposons une approche semi-supervisée pour l'extraction des relations en utilisant parallèlement les méthodes suivantes : (i)l'extraction de relations basée sur la signature des propriétés, (ii) la construction de patrons d'extraction à partir des résumés présents dans les pages de DBpedia, (iii) l'annotation manuelle d'un ensemble de textes pour définir des patrons syntaxiques pour extraire les relations. Nous avons évalué notre approche sur deux types de corpus : (i) un premier corpus est composé d'articles de journaux spécialisés, tels que aufeminin.com et Cosmétique Hebdo, (ii) un deuxième corpus est constitué d'un ensemble de phrases collectées sur sur le Web. L'évaluation présentée dans cet article combine les résultats des trois méthodes.

Mots-clés : Base de connaissances, extraction de connaissances, reconnaissance d'entités nommées, extraction de relations, RDF, ontologies, TAL, cosmétique.

1 Introduction

Avec la croissance exponentielle de la quantité de données numériques et la diversification de leurs sources, les traitements automatiques des données deviennent chaque jour un enjeu plus crucial. La mise en place de tels traitements est particulièrement difficile lorsque les données sont hétérogènes, distribuées et peu structurées comme c'est le cas pour des textes libres de pages d'information existant sur le Web. Aussi, la grande masse de données textuelles publiées sur le Web a donné lieu à de nombreux travaux en Extraction d'Information dont un but est d'enrichir ces textes tout-venant par des méta-données structurées et fortement connectées afin d'exploiter ces liens et leur sémantique dans le traitement de l'information comme la recherche, la notification, l'agrégation, etc. L'extraction d'information (EI) consiste à extraire des informations pertinentes à partir des textes telles que des entités (par exemple, *personnes*, *organisations*, *dates*), et des relations (*née à*, *racheté par*, *est du type*, ..).

Dans le cadre du projet de recherche collaboratif SMILK [1] (Social Media Intelligence and Linked Knowledge), deux questions réciproques ont été soulevées : 1) Comment utiliser le Traitement Automatique du Langage Naturel (TALN) pour contribuer au développement du Web des données liées, 2) comment utiliser le web des données liées pour contribuer à la résolution

1. https ://project.inria.fr/smilk/fr/

de tâches du TALN. Au cœur de ces problématiques, nous nous focalisons sur la tâche d'extraction de relations entre deux entités à partir des textes en français. Par exemple, dans la phrase simpliste "La Vie est Belle contient du linalol", l'objectif est d'extraire la relation "contient" entre le parfum "La Vie est Belle" et le composant "Linalol". On peut dès lors structurer la connaissance sous forme de triplets RDF. D'un point de vue applicatif, porter l'intérêt sur les relations sémantiques permet d'enrichir les méta-données afin de procéder à une recherche plus précise, moins ambiguë.

Dans cet article, nous partageons notre expérience concernant l'extraction de relations en utilisant d'une part des outils et des techniques utilisées dans le domaine du TALN, d'autre part des ontologies publiées sur le Web des données ouvertes. Nos collaborations antérieures nous ont conduites à travailler dans le secteur du Luxe et plus précisément dans la Cosmétique, c'est pourquoi nous baserons nos expériences sur l'utilisation de ProVoc, décrite par (Lopez *et al.*, 2016), qui est un vocabulaire permettant de décrire les produits sur le Web, construit à partir de scénarios majoritairement issus de la Cosmétique. ProVoc se positionne comme une extension de l'ontologie GoodRelations, décrite par (Hepp, 2008) et se concentre sur une représentation fine des entités d'intérêts (gammes de produits, composants, créateurs, *etc.*) et des relations les reliant (`belongsToBrand`, `hasComponent`, `hasDesigner`), *etc*.

Dans la suite de l'article, nous décrivons les travaux antérieurs (section 2), puis le cadre général de notre approche (section 3). Dans la section 4, nous nous focalisons sur la tâche d'extraction de relations que nous évaluons dans la section 5. La dernière section met en avant nos conclusions et nous y indiquons nos perspectives.

2 Approche générale

L'approche générale d'extraction de connaissances que nous avons suivie est illustrée en Figure 1. Le système prend en entrée un texte non structuré, transmis à un premier module de reconnaissance d'entités nommées, puis à un module d'extraction des relations.

La reconnaissance des entités doit être en mesure de reconnaître les entités qui peuvent être typées avec des classes de ProVoc dont les labels sont en anglais, *i.e.* personnes, entreprises, divisions d'une entreprise, noms de marques, noms de produits, noms de gammes de produits, composants de produits. Cette reconnaissance s'effectue par des outils complémentaires :

— Renco décrit par (Lopez *et al.*, 2014), basé sur des règles lexico-syntaxiques, qui permet d'extraire à partir des textes français des produits, leurs ingrédients, gammes de produits, marques, divisions et groupes. Les ressources utilisées par Renco étant limitées pour les noms de produits cosmétiques, nous avons construit automatiquement un lexique complémentaire contenant 1130 noms de produits à partir de pages Wikipedia français telles que "Liste de parfums" (en utilisant le service Web MediaWiki).

— Holmes Semantic Solutions [2], basé sur une approche hybride (symbolique et statistique) pour reconnaître les noms de personnes.

Une fois les entités nommées annotées, la seconde étape consiste à extraire automatiquement les relations qui existent entre ces entités, en se basant sur un analyseur syntaxique et sur la base de connaissances DBpedia. L'extraction de relations est détaillée dans la section suivante.

2. http ://www.ho2s.com/fr/

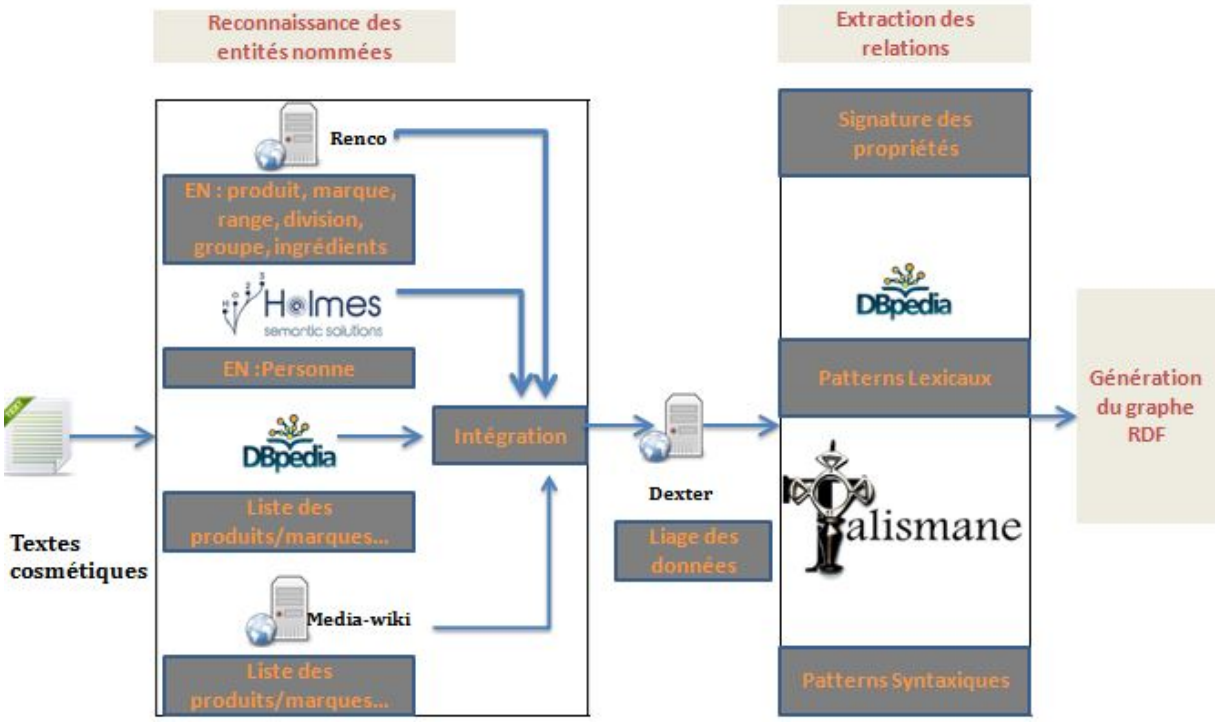


FIGURE 1 – Processus général d'extraction de connaissances à partir de textes

3 Extraction de relations

Dans cette section, nous faisons un tour d'horizon des travaux antérieurs menés sur la tâche d'extraction de relations puis nous décrivons notre méthode.

3.1 Positionnement

La tâche d'extraction des relations consiste à détecter des liens sémantiques qui existent entre différentes entités. Les différentes approches existantes peuvent être classées en trois catégories :

— Une première approche, supervisée, consiste à considérer le problème de l'extraction de relations comme un problème de classification, et à utiliser un classifieur linéaire, auquel il faut fournir un ensemble d'exemples positifs et un ensemble d'exemples négatifs pour l'apprentissage. Par exemple, les méthodes dites *feature-based* et *kernel-based* décrites dans (Nebhi, 2013) adoptent cette approche.

— Une deuxième approche, semi-supervisée, permet au système d'apprendre itérativement des patrons et des instances à partir d'un nombre réduit d'instances : les patrons sont

Propriété	Domaine	Co-domaine
`belongsToBrand`	`pv:ProductOrServiceRange` `gr:ProductOrService`	`gr:Brand`
`belongsToDivision`	`gr:Brand`	`pv:Division`
`belongsToGroup`	`pv:Division`	`gr:BusinessEntity`
`hasComponent`	`gr:ProductOrService`	`pv:Component`
`hasFragranceCreator`	`gr:ProductOrService`	`foaf:Person`
`hasRepresentative`	`rdf:Resource`	`foaf:Person`

TABLE 1 – Liste des propriétés à extraire

utilisés pour extraire des nouvelles relations à partir de l'ensemble de données, qui sont ajoutées à l'ensemble des exemples. Cette opération est répétée jusqu'à ce qu'aucune nouvelle relation ne puisse être apprise à partir de l'ensemble de données. Les systèmes DIPRE, Snowbal, supervision distance, BOA décrits par (Kumar & Manocha, 2007) adoptent cette approche.

— Une troisième approche, non supervisée, est la génération des patrons d'extraction. Elle est décrite dans (Nebhi, 2013). Cette méthode consiste à collecter des paires de mots avec les chaînes de caractères (string) qui les séparent, pour calculer la cooccurrence de termes ou pour générer les patrons. RdfLiveNews est un système décrit dans (Gerber *et al.*, 2013) qui adopte cette approche pour générer une base de connaissances RDF à partir de textes.

Dans notre travail, nous combinons deux approches pour extraire les relations : (i) une approche semi-supervisée qui consiste à extraire des relations dans les textes en utilisant les propriétés définies dans des ontologies, (ii) une approche supervisée qui consiste à extraire des relations à l'aide de patrons syntaxiques définis manuellement en analysant un corpus de textes dit d'entrainement, au préalable annoté également manuellement dans ce but.

3.2 Processus d'extraction des relations

Dans le cadre de cette étude, nous avons utilisé six propriétés définies dans le vocabulaire ProVoc, listées dans le tableau 1 avec l'indication de leurs domaines et co-domaines correspondant aux types des entités nommées concernés par ces relations. Dans les sous-sections suivantes, nous décrivons nos approches semi-supervisée et supervisée.

3.2.1 Extraction de propriétés basée sur des ontologies

La première approche d'extraction repose sur (i) les signatures des relations telles que définies dans les ontologies ProVoc et GoodRelations, (ii) des patrons lexicaux générés à partir de DBpedia.

3.2.1.1 Extraction de relations basée sur la signature des propriétés

ProVoc et GoodRelations fournissent les signatures de chaque propriété à partir desquelles des règles d'extraction sont automatiquement générées. Par exemple, la propriété `belongsTo-`

`Division` implique que le sujet doit être de type `Brand` (resp. `Division`) et que l'objet doit être du type `Division` (resp. `Brand`). La règle suivante est ainsi générée :

belongsToDivision : (sujet=Brand ET objet=Division) OU (sujet=Division ET objet=Brand)

Ces règles sont mises en oeuvre en utilisant Renco qui permet de nous fournir le typage des EN, et appliquées sur trois types de relations `belongsToBrand`, `belongsToDivision` et `belongsToGroup`. Nous avons défini une règle pour chacune des relations. L'application de ces règles sur le texte annoté par les outils de reconnaissance d'entités nommées revient à projeter les signatures de propriétés sur le texte pour repérer les entités ayant pour types respectifs le domaine et co-domaine d'une propriété recherchée. L'existence d'une phrase contenant deux types différents d'entités nommées implique que nous avons forcément une relation entre ces deux entités nommées.

3.2.1.2 Extraction des relations basée sur des patrons lexicaux à partir de DBpedia

En nous inspirant de l'approche proposée par (Gerber & Ngomo, 2011) nous avons construit des patrons lexicaux (sans utilisation de la syntaxe) à partir de l'analyse des résumés des produits cosmétiques décrits dans DBpedia en l'interrogeant avec des requêtes SPARQL. Nous avons écrit une requête SPARQL pour interroger la page DBpedia dont le nom est "Liste de parfums", le but étant d'extraire des informations concernant les parfums tels que leurs noms, leurs marques et le résumé présenté pour chaque parfum.

```
select  ?parfum_name ?brand_name ?abstract
where {
       ?parfum_list rdfs:label "Liste de parfums" @fr.
       ?parfum_list dbpedia-owl:wikiPageWikiLink ?parfum.
       ?parfum rdfs:label ?parfum_name.
       ?parfum prop-fr:marque ?brand.
       ?brand rdfs:label ?brand_name.
       ?parfum dbpedia-owl:abstract ?abstract.
       FILTER (LangMatch (lang(?abstract), "fr" ))
       FILTER (LangMatches (lang(?parfum_name), "fr"))
       FILTER (LangMatches (lang(?brand_name), "fr"))
 }
```

A partir de ces trois principales informations, nous avons considéré les patrons lexicaux comme les éléments textuels se situant entre deux entités respectant la signature d'une propriété donnée. En analysant par exemple les informations relatives au parfum "Allure Homme", nous obtenons que ce parfum appartient à la marque "Chanel" et que son résumé est "Allure Homme est un parfum masculin de Chanel, créé par Jacques Polge et sorti en 1999". Nous avons considéré que les éléments lexicaux qui se trouvent entre le nom de parfum et le nom de la marque représentent le patron lexical « **est un parfum masculin de** » . Nous avons obtenu neuf patrons lexicaux.

3.2.2 Extraction de relations basée sur des patrons syntaxiques

En complément des approches lexicales présentés ci-avant, nous avons développé une approche basée sur les relations de dépendances syntaxiques ce qui a pour avantage de ne pas s'en tenir exclusivement aux éléments lexicaux se situant entre les entités. En particulier, les dépendances syntaxiques entre le verbe, le sujet et l'objet permettent d'assurer la cohésion de ces trois éléments. De même, la présence d'une préposition dans un syntagme déterminatif, comme dans le cas de "Coco Noir de Chanel" permet d'identifier une relation d'appartenance (*belongsTo*). Dans le cadre de cette étude, nous avons utilisé l'analyseur syntaxique Talismane décrit dans (Urieli, 2013), un des rares analyseurs pour le français sous licence GPL qui permette le repérage et l'étiquetage des dépendances syntaxiques entre les mots.

Afin de développer les patrons syntaxiques, nous avons construit un jeu d'apprentissage constitué de 58 phrases issues de différents journaux tels que Cosmétique Hebdo, Cosmétique Mag. Ce corpus contient 55 relations de type `hasComponent`, 27 relations de type `hasFragranceCreator` et 24 relations de type `hasRepresentative`. Une étape de pré-traitement a consisté à annoter les résultats issus de la reconnaissance des entités nommées (cf. Section 2) de telle façon à ce que Talismane en tienne compte. Le type des EN reconnues est également indiqué à Talismane en utilisant l'attribut `comment` de son API.

A partir du jeu d'apprentissage nous avons défini manuellement 30 patrons pour extraire la relation `hasRepresentative`, 22 patrons pour extraire la relation `hasFragranceCreator` et 63 patrons pour extraire la relation `hasComponent`. Considérons par exemple la phrase suivante : « La ligne s'anime également en mai avec une édition limitée Summer, rafraîchie d'ananas, création d'Anne Flipo et de Carlos Benam (IFF), dans un flacon orange (EdT vapo 50 et 100 ml, 42 et 48) ». Cette phrase contient une relation de type `hasComponent` entre `Summer` et `ananas` et deux relations de type `hasFragranceCreator` entre `Summer` et `Anne Flipo` et entre `Summer` et `Carlos Benam`.

Concrètement, pour extraire la relation `hasComponent` et son objet, nous avons écrit la règle syntaxique 1. Cette règle consiste à extraire un token dont la partie du discours (POS) est un participe passé (VPP) ayant une relation de type modificateur ou dépendance, suivi d'un token dont le POS est "préposition et déterminant" ou seulement préposition ayant une relation de dépendance de type syntagme prépositionnel, suivi d'un token de type nom propre ayant une relation de préposition avec son prédécesseur et qui représente une entité nommée de type "Component". Pour extraire la relation `hasFragranceCreator` et son objet, nous avons écrit les règles syntaxiques 2 et 3. L'extraction du sujet attachée à une propriété est effectuée en suivant le même principe : pour identifier le sujet, nous avons implémenté 22 patrons. Par exemple, pour extraire le sujet des propriétés extraites de la phrase ci-dessous, nous avons défini la règle 4.

1 "SI depRel(,VPP)=(mod ou dep) ET depRel(VPP,(P+D ou P))=de obj ET depRel((P+D ou P),NPP)=prep ET type (NPP)=ingredient".
2 "SI depRel(0,VPP)=mod ET depRel(VPP,NC)=(mod ou suj ou prep) ET depRel(NC,(P ou P+D))=dep ET depRel((P ou P+D),NPP)=prep ET type (NPP)=PER
alors (NPP)=FragranceCreator"
3 "SI depRel(0,VPP)=mod ET depRel(VPP,NC)=(mod ou suj ou prep) ET depRel(NC,(P ou P+D))=dep ET depRel((P ou P+D),NPP)=prep ET type NPP =PER + isCoordination"
4 "Si depRel(0,V)=root ET depRel(V,P)=mod ET depRel(P,NC)=(prep ou obj) ET de-

pRel(NC,NPP)=mod ET type(NPP)=product"

Le résultat du processus est un ensemble de triplets RDF décrivant d'une part chaque entité reconnue avec son type et son label, et d'autre part chaque relation extraite avec son sujet et son objet. Par exemple :

```
smilk:Txt1 schema:about skp:Summer ; rdfs:comment "....." .

skp:Summer a gr:ProductOrService ; rdfs:label "Summer" ;
           pv:hasComponent skc:ananas ;
           pv: hasFragranceCreator skf:Anne_Flipo ;
           pv: hasFragranceCreator skf:Carlos_Benam .

skc:ananas a pv:Component ; rdfs:label  "ananas" .

skf:Anne_Flipo a foaf:Person ; rdfs:label  "Anne Flipo" .

skf:Carlos_Benam a foaf:Person ; rdfs:label "Carlos Benam" .
```

4 Mise en oeuvre et évaluation

4.1 Corpus et base de connaissance

Nous avons construit deux corpus pour tester notre approche :

— Le **corpus L'Oréal** est issu de l'outil Factiva et contient des articles journalistiques contenant la mention L'Oréal. Ce corpus est composé de 392 phrases issues de différents journaux tels que aufeminin.com, Cosmétique Hebdo, Cosmétique Mag. Ces phrases peuvent contenir ou non des relations ProVoc. Ce corpus contient 84 relations de type `belongsToBrand`, 79 relations de type `hasComponent`, 38 relations de type `hasFragranceCreator` et 44 de type `hasRepresentative`.

— Le **corpus Web** est issu d'une extraction manuelle de phrases issues du Web par le biais du moteur de recherche Google. Nous avons cherché le nom d'un parfum dans Google, par la suite nous avons choisi les dix premiers liens après les annonces tels que des liens vers des sites commerciaux ou des blogs cosmétiques. Ce corpus est composé de 119 phrases. Il contient 69 relations de type `hasFragranceCreator`, 62 relations de type `hasComponent`, 81 relations de type `hasRepresentative` et 43 relations de type `belongsToBrand`.

Notre système a permis de générer 325 triplets à partir du *corpus L'Oréal*, et 988 triplets à partir du *corpus Web*. La pertinence de ces triplets est évaluée dans les sections suivantes.

4.2 Évaluation de l'extraction des entités nommées

Nous avons réalisé une première évaluation de l'extraction des EN, nous avons considéré un sous-ensemble du *corpus du Web* composé de 25 phrases et nous avons cherché à évaluer l'extraction des entités nommées de type Product, Brand, Personne et Component. Les phrases contiennent 31 entités de type Product, 19 entités de type Brand et 20 entités de type Component.

Type EN	Rappel	Précision	Outils
Product	0.53	0.89	Renco
	0.66	0.91	Renco + Wikipedia
	0.17	0.5	DBpedia Spotlight
Brand	0.47	0.9	Renco
	0.73	0.91	Renco + Wikipedia
	0.61	0.84	DBpedia Spotlight
Component	0.55	0.73	Renco
	0.68	0.92	DBpedia Spotlight

TABLE 2 – Évaluation de l'extraction des entités nommées

Nous avons constaté que Holmes permet d'extraire les entités nommées de type Personne avec une valeur de rappel et de précision de 1. Pour les autres types d'EN, nous avons comparé la précision et le rappel de l'extraction d'EN avec Renco, avec Renco et les listes d'EN extraites de Wikipedia et avec DBpedia Spotlight [3], un outil de référence dans l'état de l'art de l'annotation sémantique de textes, qui repose sur DBpedia. Le tableau 2 présente les résultats obtenus. Sans surprise, pour l'extraction d'EN de type Product ou Brand, le rappel et la précision obtenus en complémentant Renco d'une liste d'EN extraite de Wikipedia sont meilleurs qu'en utilisant Renco seul. La valeur moyenne du rappel obtenue peut être expliquée par l'absence de certaines EN dans les pages Wikipedia exploitées et une tokenisation incorrecte de certaines EN par Renco. Par exemple, l'EN "La vie est belle" est considérée par Renco comme quatre tokens au lieu d'un seul. Comparés aux résultats obtenus avec DBpedia Spotlight, notre approche obtient de meilleures valeurs de rappel et de précision. Cependant, pour les EN de type Component, pour lesquelles une liste d'EN n'a pas été extraite de Wikipedia, DBpedia Spotlight obtient de meilleurs résultats.

4.3 Évaluation de l'extraction des relations

Le protocole de l'évaluation consiste à faire des expérimentations sur les deux corpus et dans le cas du corpus construit à partir de textes du Web à raffiner l'évaluation en considérant un sous-ensemble où la détection des entités nommées a été manuellement validée et au besoin corrigée ou complétée.

4.3.1 Évaluation de l'extraction de relations sur le corpus des articles du journaux

Le tableau 3 montre les résultats obtenus sur le corpus des articles du journaux. Plusieurs raisons peuvent expliquer la faiblesse de ces résultats :

— L'impossibilité de définir toutes les règles syntaxiques qui peuvent exprimer les relations à extraire.
— L'utilisation de règles syntaxiques seulement et l'absence de règles lexicales peuvent limiter l'approche de l'extraction des relations en diminuant la valeur de la précision. Pour

3. http ://demo.dbpedia-spotlight.org/

Propriétés	Méthode d'extraction	Rappel	Précision
belongsToBrand	signature des propriétés	0.17	0.46
hasComponent	Patrons syntaxiques	0.06	0.25
hasFragranceCreator	Patrons syntaxiques	0.21	0.20
hasRepresentative	Patrons syntaxiques	0.11	0.18

TABLE 3 – Évaluation de l'extraction de relations sur les articles de journaux spécialisés

Propriétés	Méthode d'extraction	Rappel	Precision
belongsToBrand	signature des propriétés	0.4	0.4
hasComponent	Patrons syntaxiques	0.2	1
hasFragranceCreator	Patrons syntaxiques	0.43	0.45
hasRepresentative	Patrons syntaxiques	0.3	0.52

TABLE 4 – Évaluation de l'extraction de relations sur un corpus d'articles sélectionnés sur le Web

certaines relations, une même règle syntaxique peut conduire à extraire deux relations différentes, comme par exemple `hasFragranceCreator` et `hasRepresentative`. Nous citons par exemple la phrase suivante « Monica Bellucci reste l'égérie de Rouge Dior sous l'objectif de Tyen. ». En réalité, cette phrase représente la relation `hasRepresentative` entre le produit Rouge Dior et l'égérie « Monica Belluci », alors que notre approche permet d'extraire en plus une relation de type `hasFragranceCreator` entre les mêmes entités.

— La non détection de certaines EN. Dans notre approche, nous utilisons DBpedia et Wikipedia pour détecter les entités nommées de type produit ou marque dans le cas où Renco ne peut pas les identifier, mais Wikipedia et DBpedia peuvent ne pas contenir toutes les marques et noms du parfums. Par exemple les parfums « Spicebomb » et « the essence » n'y figurent pas. De plus, l'expression du nom d'un produit différemment dans le texte et dans Wikipedia ou DBpedia empêche la détection. Par exemple, un même produit est représenté par « 1 Million » dans le corpus de textes et par « one Million » dans Wikipedia dont le token "one" est reconnu comme un nom propre et pas un nombre.

4.3.2 Évaluation de l'extraction de relations à partir d'articles sélectionnés sur le Web

Le tableau 4 présente les résultats de l'extraction des relations sur le corpus construit en sélectionnant des articles sur le Web. Les mêmes raisons citées dans l'évaluation des articles des journaux peuvent aussi expliquer les résultats de cette évaluation. De plus, nous avons supposé dans notre approche qu'une règle syntaxique doit impérativement contenir un verbe, alors que ce corpus contient des phrases sans verbe comme par exemple « Theo James, nouvel ambassadeur des parfums Hugo Boss. »

Afin d'évaluer la qualité de l'extraction de relations indépendamment de la qualité de la reconnaissance d'EN, nous avons considéré 45 phrases du corpus collecté sur le Web avec Google,

Propriétés	Méthode d'extraction	Rappel	Précision
belongsToBrand	signature des propriétés	0.625	0.5
hasFragranceCreator	Patrons syntaxiques	0.57	0.31
hasRepresentative	Patrons syntaxiques	0.32	0.69

TABLE 5 – Évaluation de l'extraction de relations sur un sous-ensemble du corpus d'articles sélectionnés sur le Web où les EN sont bien reconnues

pour lesquelles la reconnaissance des EN était correcte. Le tableau 5 présente les résultats de cette évaluation.

4.4 Synthèse

Le rappel et la précision sur le corpus des textes du Web sélectionnées avec Google sont meilleurs que ceux correspondants sur le corpus des articles de journaux. Cela peut être expliqué par la richesse des structures grammaticales dans les articles des journaux comparativement aux phrases collectées sur le Web qui expriment en général une représentation simple du produit avec donc une forme syntaxique plus ou moins régulière. Par exemple, dans le corpus du Web, la relation `hasComponent` est généralement exprimée avec un lien direct entre le parfum et le composant comme la phrase « Gentelmen Only Absolute contient notamment de la canelle, du safran et de la muscade ». Alors que dans les articles des journaux, cette relation est exprimée d'une manière plus compliquée comme « Alors que l'Agence nationale de sécurité du médicament et des produits de santé déconseille pour les enfants les produits contenant plus de 4 g de phénoxyéthanol par kilo et que la recommandation du Comité scientifique pour la sécurité des consommateurs est de ne pas dépasser 2,48 g de propylparaben par kilo, l'UFC relève des taux de 2,68 g de propylparaben et de 8,02 g de phénoxyéthanol dans le gel douche Nivea Water Lily Oil »

Le rappel et la précision de la reconnaissance des relations sur le sous-ensemble du corpus extrait du Web où la reconnaissance des entités nommées a été validée sont significativement supérieurs à ceux obtenus sur l'ensemble du corpus avec une reconnaissance d'EN automatique. En d'autres termes, la qualité de la reconnaissance des EN influe significativement sur celle de l'extraction des relations et donc des performances de l'approche dans son ensemble.

5 Conclusion

Nous avons proposé dans cet article une approche d'extraction de relations à partir de textes dans le domaine de la cosmétique, qui combine une approche semi-supervisée utilisant les signatures des propriétés déclarées dans l'ontologie du domaine Provoc, avec une approche supervisée utilisant des patrons syntaxiques définis manuellement en analysant un corpus de textes préalablement annoté manuellement. Le résultat est la production d'une base de connaissances formalisée en RDF. Nous avons évalué notre approche sur deux corpus différents, un corpus d'articles de journaux spécialisés dans le domaine de la cosmétique, similaire au corpus analysé pour produire manuellement les règles syntaxiques d'extraction, et un corpus de textes du Web sélectionnées avec Google.

Les résultats obtenus montrent bien l'intérêt d'une approche mixte dès lors qu'un domaine particulier est ciblé, et la corrélation qui existe entre qualité de l'extraction des propriétés et qualité de la reconnaissance d'entités nommées. Les pistes naturelles d'amélioration de cette approche sont d'une part l'amélioration de la reconnaissance d'entités nommées et d'autre part l'ajout d'une dimension lexicale et sémantique aux patrons syntaxiques. Une autre perspective de ce travail est de comparer les résultats ainsi obtenus avec ceux que l'on obtiendrait avec des outils d'apprentissage automatique statistique.

Références

GERBER D., HELLMANN S., BUHMANN L., SORU T., USBECK R. & NGOMO A.-C. N. (2013). Real-time rdf extraction from unstructured data streams. In *International Semantic Web Conference*, p. 135–150 : Springer.

GERBER D. & NGOMO A.-C. N. (2011). Bootstrapping the linked data web. In *1st Workshop on Web Scale Knowledge Extraction@ ISWC*, volume 2011.

HEPP M. (2008). Goodrelations : An ontology for describing products and services offers on the web. In *International Conference on Knowledge Engineering and Knowledge Management*, p. 329–346 : Springer.

KUMAR K. & MANOCHA S. (2007). Constructing knowledge graph from unstructured text. *Self*, **3**, 4.

LOPEZ C., NOORALAHZADEH F., CABRIO E., SEGOND F. & GANDON F. (2016). Provoc : une ontologie pour d'écrire des produits sur le web. In *IC2016 : 27es Journees francophones d'Ingenierie des Connaissances*.

LOPEZ C., SEGOND F., HONDERMARCK O., CURTONI P. & DINI L. (2014). Generating a resource for products and brandnames recognition. application to the cosmetic domain. In *LREC*, p. 2559–2564.

NEBHI K. (2013). A rule-based relation extraction system using dbpedia and syntactic parsing. In *Proceedings of the 2013th International Conference on NLP & DBpedia-Volume 1064*, p. 74–79 : CEUR-WS. org.

URIELI A. (2013). *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. PhD thesis, Université Toulouse le Mirail-Toulouse II.

Une approche hybride pour la détection d'influenceurs dans les médias sociaux

Namrata Patel[1], Cédric Lopez[1], Ioannis Partalas[1], Frédérique Segond[1]

VISEO TECHNOLOGIES, CENTRE R & D
4 av. Doyen Louis Weil, 38000, Grenoble, France
[nom].[prenom]@viseo.com

Résumé : Un influenceur est une personne ayant les capacités d'imposer des attitudes, des comportements, des opinions à un public. La détection de tels individus est d'un grand intérêt, par exemple pour les campagnes commerciales qui cherchent à identifier les acteurs les plus influents pour promouvoir une marque ou un produit. Dans cet article, nous nous focalisons sur la détection d'influenceurs dans les réseaux sociaux qui s'appuie généralement sur une structure de graphe dont les nœuds sont les utilisateurs et les arcs orientés sont leurs interactions. Au-delà de la structure du graphe, nous expérimentons l'impact du contenu textuel des messages sur l'influence. En considérant des critères linguistiques hypothétiques (tels que l'argumentation d'un utilisateur, l'accord/désaccord entre utilisateurs) en plus des critères numériques classiques (nombre de réponses, taille du message, nombre de relations, etc.), nous montrons par le biais d'un système d'apprentissage que certains critères linguistiques sont pertinents pour la tâche de détection d'influenceurs.

Mots-clés : Détection d'influenceurs, traitement automatique du langage naturel, système d'apprentissage

1 Introduction

"Dans un contexte marketing et dans son sens le plus large, un influenceur est un individu qui par son statut, sa position ou son exposition médiatique peut influencer les comportements de consommation dans un univers donné." [1] Un influenceur a les capacités d'imposer des attitudes, des comportements, des opinions à un public, ce qui en fait une source de communication pertinente pour les organismes désireux de promouvoir une marque, un produit, un service, ou une idée.

Depuis quelques années, poussé par les intérêts du marketing, des travaux cherchent à évaluer de façon automatique la force d'influence des individus impliqués dans les médias sociaux. La détection d'influenceurs s'appuie généralement sur une structure de graphe représentant les utilisateurs et leurs interactions pour calculer une mesure de centralité (Bonacich, 1987). Récemment, cette tâche s'est intéressée au contenu textuel généré par les utilisateurs en considérant des critères linguistiques (Kien-Weng Tan *et al.*, 2011) (Rosenthal, 2015).

Dans cet article, nous faisons l'hypothèse que les critères linguistiques (tels que l'argumentation d'un utilisateur, l'accord/désaccord entre utilisateurs) sont pertinents pour la tâche de détection d'influenceurs dans les médias sociaux. Pour ce faire, les informations sont extraites du contenu textuel par des règles linguistiques puis sont intégrées dans un système d'apprentissage automatique. Le système résultant est un système "doublement hybride" puisqu'il s'appuie sur des méthodes symboliques et statistiques d'une part, et sur la structure et le contenu textuel du réseau d'autre part.

L'article est organisé de la façon suivante : la section 2 fait un tour d'horizon sur la détection d'influenceurs dans les médias sociaux. La section suivante détaille l'architecture de notre

1. http ://www.definitions-marketing.com/definition/influenceur/

système évalué à la section 4.

2 État de l'art

Dans la littérature, la tâche de détection d'influenceurs exploite :

— les interactions entre utilisateurs en se fondant sur la théorie des graphes. Étant donné un réseau social, l'influence est calculée en analysant les informations structurelles des interactions entre utilisateurs.

— le contenu textuel des messages publiés par les individus. Le texte est analysé en se fondant sur des indicateurs linguistiques.

Les études basées sur la théorie des graphes exploitent la pléthore de mesures disponibles dans la littérature pour analyser l'information structurelle du réseau d'interactions des utilisateurs. Pour la détection d'influence en particulier, les *mesures de centralité* sont utilisées afin d'identifier les nœuds les plus importants d'un réseau (Bonacich, 1987). Parmi ces mesures de centralité, une mesure indicative est la *centralité d'interférence*, qui exprime le nombre de fois qu'un nœud donné est dans le chemin le plus court entre deux nœuds quelconques dans le réseau. Dans la même famille, *PageRank* est l'une des mesures les plus connues pour le classement des nœuds (Page *et al.*, 1999). Par ailleurs, Kempe *et al.* (2003) utilisent des *modèles de propagation* dans le but de spécifier comment les actions (par exemple les *retweets* d'un message dans Twitter) se propagent à travers le réseau social et Reid & Ng (2000) montrent qu'il y a une forte corrélation entre les tours de conversation et l'influence.

Les études récentes fondées sur la prise en compte du contenu textuel cherchent à identifier les traits de comportement influents par le biais de marqueurs linguistiques présents dans les messages. Biran *et al.* (2012) et (Rosenthal, 2015) proposent plusieurs marqueurs tels que la persuasion, l'accord et le désaccord, la structure de dialogue et les sentiments, et proposent une approche d'apprentissage automatique pour détecter les influenceurs. D'autres travaux se focalisent sur les opinions véhiculées par les messages (Bigonha *et al.*, 2012). Cette lignée de travaux fournit des résultats prometteurs de par une analyse fine des comportements influents.

Enfin, il existe quelques rares travaux cherchant à combiner l'analyse structurelle et l'analyse du contenu. Par exemple, Weng *et al.* (2010) ajoutent un biais à la mesure de PageRank dans le calcul d'influence entre utilisateurs, en tenant compte du sujet abordé dans les messages ; Katsimpras *et al.* (2015) proposent une approche supervisée afin de détecter des nœuds influents pour un sujet donné ; le contenu du message n'est exploité ici que sous l'angle du sujet abordé.

Dans les travaux antérieurs, la détection d'influenceurs se fait par apprentissage automatique. Notre approche intègre un système à base de règles permettant l'analyse du contenu des messages selon plusieurs descripteurs que nous expérimentons.

3 Description de l'approche hybride

Notre approche se compose de quatre phases successives :

1. Constitution du jeu de données
2. Analyse linguistique
3. Génération d'un modèle d'apprentissage

4. Calcul de scores d'influence (par message et par auteur) et classement

Chacune de ces phases est décrite dans la suite.

3.1 Constitution des jeux de données

Notre jeu de données initial est constitué d'un ensemble de fils de discussions en anglais extraits d'un forum de cosmétique [2]. Nous avons recueilli plus de 5000 fils de discussions et les avons divisé aléatoirement en trois parties : 1000 fils sont dédiés à l'analyse et au développement de règles linguistiques, 1000 fils sont dédiés à l'entraînement du modèle d'apprentissage, et 3000 fils sont réservés à l'évaluation de notre approche. Chacun des 18085 messages constituant les 1000 fils dédiés à l'entraînement du modèle d'apprentissage ont été annotés manuellement. Il s'agissait de décider de façon booléenne si un message a un pouvoir d'influence ou non, selon ses propres critères personnels.

3.2 Analyse linguistique

La phase d'identification des critères s'est appuyée sur les travaux cités dans la section précédente et sont présentés dans le tableau 1. Les critères sont catégorisés en tant que "linguistique" ou "non linguistique". Les critères non linguistiques sont calculés par des fonctions de comptage ou booléennes. Les critères linguistiques sont calculés à base de règles linguistiques.

Critère	**Catégorie**	**Nature**	**Sortie (type)**
isFirstPost ?	non-linguistique	Position d'un message dans un fil	booléen
isSecondPost ?	non-linguistique	Position d'un message dans un fil	booléen
isPenultimateost ?	non-linguistique	Position d'un message dans un fil	booléen
isLatestPost ?	non-linguistique	Position d'un message dans un fil	booléen
sizeOfMessage	non-linguistique	Information quantitative	$0 < x < n$
RegistrationDate	non-linguistique	Date	date
Location of the user	non-linguistique	Emplacement	string
Elongation	linguistique	Style d'écriture	booléen
Uppercase	linguistique	Style d'écriture	booléen
Exclamation	linguistique	Style d'écriture	booléen
Interrogation	linguistique	Style d'écriture	booléen
Nb of premises	linguistique	Argumentation	$0 < x < n$
conclusion ?	linguistique	Argumentation	booléen
ArgumentInFirstSentence	linguistique	Argumentation	booléen
Advising	linguistique	Argumentation	$0 < x < n$
Advising	linguistique	Accord	booléen
Advising	linguistique	Désaccord	booléen

TABLE 1 – Description des caractéristiques extraites qui servent d'entrée pour le modèle d'apprentissage

A chaque critère linguistique est associé un module de règles que nous avons développé manuellement à partir du jeu de données dédié (3.1). Chaque règle s'appuie sur une analyse morphosyntaxique fournie par Holmes Semantic Solutions [3].

Les textes sont ainsi automatiquement annotés selon ces critères qui serviront par la suite d'entrée pour le modèle d'apprentissage supervisé.

2. Nous en dissimulons le nom pour des raisons de confidentialité
3. http ://www.ho2s.com/fr/

3.3 Génération d'un modèle d'apprentissage

Suite à l'extraction des valeurs de chaque critère, l'ensemble de messages annotés est représenté sous forme matricielle : chaque ligne représente un message et chaque colonne représente une caractéristique. La matrice est remplie en fonction des valeurs des critères extraits par message et sert comme telle d'entrée pour le modèle d'apprentissage supervisé. Nous avons choisi les *Forêts d'arbres décisionnels* qui sont robustes et très utilisés dans une variété d'applications

3.4 Calcul du score d'influence

Notre modèle génère un score d'influence par message, et représente la probabilité de répondre positivement à chaque critère. Ces scores d'influence sont ensuite agrégés afin de produire un score d'influence final par auteur de message. Cette agrégation se fait en exploitant l'information structurelle présente dans le réseau d'interactions des utilisateurs (auteurs).

Soit $U = \{u_1, u_2, \ldots, u_n\}$ l'ensemble d'utilisateurs dans un réseau social et $S_u = \{s_1, s_2, \ldots, s_{K_u}\}$ l'ensemble de scores par message d'un utilisateur donné u, où $K_u =$ nombre de messages postés par l'utilisateur. Le score final par utilisateur est alors défini par :

$$Inf(u) = \frac{\frac{1}{K_u}\sum_{i=1}^{K_u} s_i}{\max_{u'} \frac{1}{K_{u'}}\sum_{j=1}^{K_{u'}} s_j}$$

4 Evaluation

4.1 Expérimentation

Nous avons entraîné notre modèle de forêts d'arbres décisionnels en effectuant une recherche aléatoire couplée avec une validation croisée à 5-plis afin de régler ses paramètres : (1) nombre d'arbres $\in [50, 500]$, (2) profondeur $\in [2, 10]$ et (3) critère d'information $\in \{entropie, gini\}$. Nous avons entraîné deux versions du modèle, avec et sans critères linguistiques afin d'évaluer leur pertinence. Les deux modèles ont été optimisés pour ROC-AUC qui est une mesure de la probabilité qu'une instance positive soit plus haute dans le classement qu'une instance négative.

4.2 Résultats

La figure 1 représente les courbes ROC-AUC pour les deux modèles. On remarque qu'avec une tolérance de 30% de faux positifs, le système peut obtenir jusqu'à 82% de vrais positifs en considérant les caractéristiques linguistiques. Pour mieux évaluer la pertinence de ces caractéristiques, nous avons également calculé le classement de l'ensemble des caractéristiques par ordre d'importance pour le modèle (*cf.* Figure 2).

On remarque que la caractéristique la plus pertinente est la taille du message, qui reflète naturellement le fait que les messages longs contiennent plus d'instances de caractéristiques linguistiques que des messages courts. Ce qui est particulièrement remarquable est le classement des critères d'argumentation (plus précisément des prémisses) et d'élongation lexicale (par exemple "ce produit est suuuuuuuper"). Ainsi, entre deux critères non linguistiques (numériques) traditionnellement utilisés dans la tâche de détection d'influenceurs, se positionnent les critères liés à l'argumentation et au conseil, ce qui dessine nos perspectives à court terme.

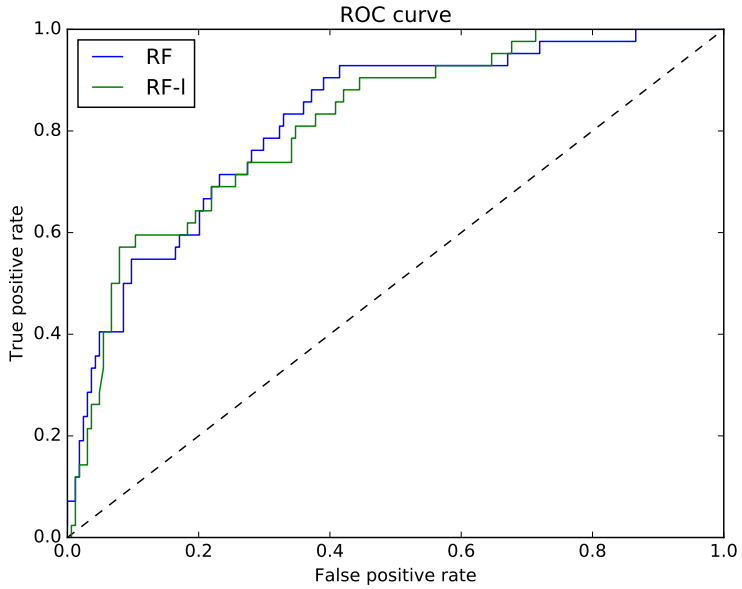


FIGURE 1 – Courbes ROC-AUC pour les deux modèles.

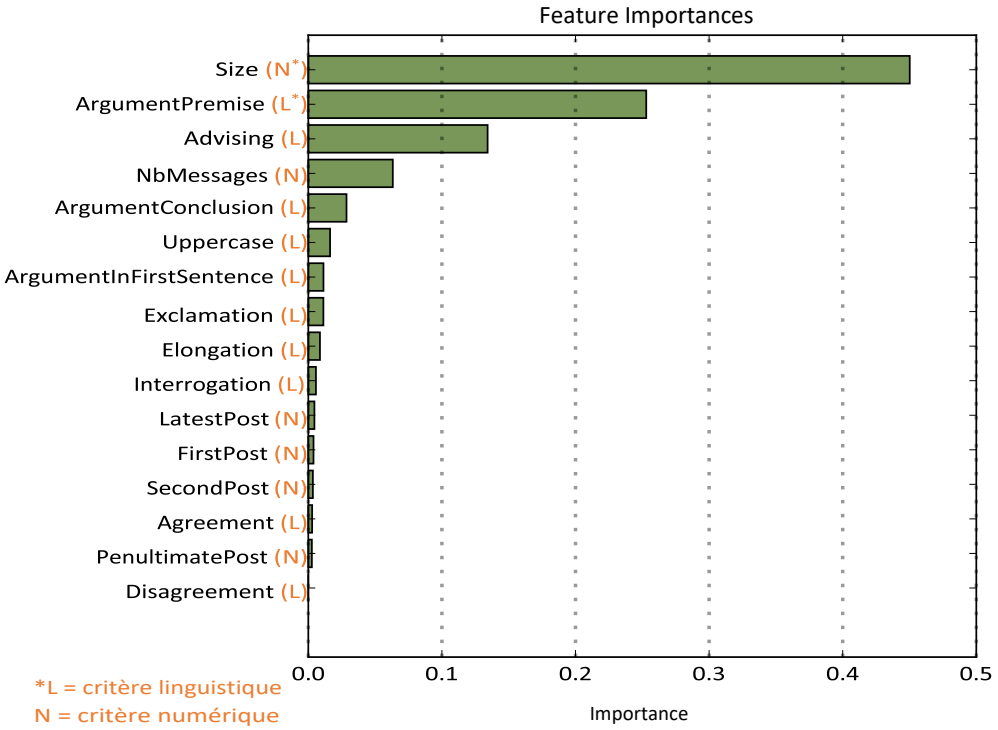


FIGURE 2 – Caractéristiques par ordre décroissant de pertinence pour le modèle.

5 Conclusion

Dans cet article, nous avons présenté une approche hybride pour la détection d'influenceurs qui s'appuie sur des méthodes symboliques et statistiques d'une part, et sur la structure et le contenu textuel des réseaux d'autre part. Ayant confronté des critères linguistiques (tels que l'argumentation d'un utilisateur, l'accord/désaccord entre utilisateurs) à des critères numériques classiques (nombre de réponses, taille du message, nombre de relations, etc.) vis à vis de la tâche de détection d'influenceurs, nos résultats confirment la pertinence des premiers dans la détection d'influence. Plus précisément, l'argumentation et la présence d'élongation apparaissent parmi les critères les plus pertinents.

Motivés par ces résultats, nous chercherons à améliorer nos modules linguistiques par une analyse structurelle du discours. Nous explorerons également une analyse basée sur la théorie des graphes afin de mieux exploiter l'information structurelle présente dans le réseau d'interactions des utilisateurs.

6 Remerciement

Ce travail de recherche est soutenu par la commission européenne Eurostars dans le cadre du projet SOMA (E9202).

Références

BIGONHA C., CARDOSO T. N., MORO M. M., GONÇALVES M. A. & ALMEIDA V. A. (2012). Sentiment-based influence detection on twitter. *Journal of the Brazilian Computer Society*, **18**(3), 169–183.

BIRAN O., ROSENTHAL S., ANDREAS J., MCKEOWN K. & RAMBOW O. (2012). Detecting influencers in written online conversations. In *Proceedings of the Second Workshop on Language in Social Media*, p. 37–45 : Association for Computational Linguistics.

BONACICH P. (1987). Power and Centrality : A Family of Measures. *American Journal of Sociology*, **92**(5), 1170–1182.

KATSIMPRAS G., VOGIATZIS D. & PALIOURAS G. (2015). Determining influential users with supervised random walks. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, p. 787–792, New York, NY, USA : ACM.

KEMPE D., KLEINBERG J. & TARDOS E. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, p. 137–146 : ACM.

KIEN-WENG TAN L., NA J.-C. & THENG Y.-L. (2011). Influence detection between blog posts through blog features, content analysis, and community identity. *Online Information Review*, **35**(3), 425–442.

PAGE L., BRIN S., MOTWANI R. & WINOGRAD T. (1999). *The PageRank Citation Ranking : Bringing Order to the Web.* Technical Report 1999-66, Stanford InfoLab.

REID S. A. & NG S. H. (2000). Conversation as a resource for influence : Evidence for prototypical arguments and social identification processes. *European Journal of Social Psychology*, **30**(1), 83–100.

ROSENTHAL S. (2015). *Detecting Influencers in Social Media Discussions*. PhD thesis, Columbia University.

WENG J., LIM E.-P., JIANG J. & HE Q. (2010). Twitterrank : Finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, p. 261–270 : ACM.

L'ontologie PHO en Histoire des Sciences et Techniques

Bruno Rohou[1,2], Sylvain Laubé[1], Serge Garlatti[2]

[1] CENTRE FRANÇOIS VIÈTE (EA 1161), Université Bretagne Occidentale, Brest, France
bruno.rohou@univ-brest.fr

[2] sylvain.laubé@univ-brest.fr

[3] IMT ATLANTIQUE, LAB-STICC Univ. Bretagne Loire, F-29238 Brest,France
serge.garlatti@imt-atlantique.fr

Résumé : Cet article présente un modèle en SHS de périodisation des ports (HST-PORT) et une ontologie de référence, appelée Port History Ontology (PHO) dont les classes, les propriétés et la structure sont issues du modèle HST-PORT et des sources historiques disponibles. Cette ontologie réutilise l'ontologie CIDOC CRM et DOLCE. Comme CIDOC CRM ne permet la représentation du temps (instant et intervalle) que par des chaînes de caractères, nous avons réutilisé l'ontologies "OWL-Time" pour définir les périodes de stabilité (intervalles de temps). Afin d'assurer la géolocaliation des artefacts, nous avons réutilisé l'ontologie WGS84-pos.

Mots-clés : Ontologies, Histoire des sciences et techniques, périodisation des ports, CIDOC-CRM, OWL-TIME.

1 Introduction : positionnements et contextes

Notre travail de recherche s'insère dans le programme du Centre F. Viète "Histoire comparée des paysages culturels portuaires" et porte sur la compréhension de l'évolution scientifique et technologique des ports de Brest (France), Mar del Plata et Rosario en Argentine à l'époque contemporaine [1]. L'hypothèse de recherche est de considérer un port comme un macro-système technologique complexe (Hughes *et al.*, 1987) dont l'évolution spatio-temporelle en tant qu'artefact s'inscrit dans une histoire des sciences et des techniques. Cette évolution spatio-temporelle peut être considérée comme multi-échelle tant sur l'espace que le temps. Le port est lui-même constitué d'artefacts en interaction (à différents niveaux de granularité) comme des formes de radoub, des jetées, des quais, des grues, des amarrages, ou encore des unités de productions industrielles (forges, corderies, etc.). Ces artefacts sont considérés comme indicateurs signifiants de cette évolution. Périodiser et comparer les ports nécessite de caractériser des moments de ruptures liés aux évolutions de ces artefacts et d'identifier des périodes plus ou moins longues où le système est stable entre deux ruptures. Dans le champ des SHS, nous avons développé un modéle d'évolution des ports, appelé HST-PORT. Ce modèle considère l'activité humaine caractérisée par ses rapports entre artefacts, acteurs et savoirs. [2]

Notre objectif est de bâtir une histoire comparée sur un grand nombre de ports, ce qui implique de construire et de valider de nouvelles méthodes de travail en humanités numériques. Il s'agit de concevoir de nouveaux systèmes d'information fondés sur des ontologies et le web sémantique (Bruneau *et al.*, 2015). Ces systèmes ont notamment pour objet d'indexer, de publier et d'interroger des sources historiques afin de produire cette histoire comparée. La modélisation

1. voir http://brmdp.hypotheses.org/
2. voir http://brmdp.hypotheses.org/269

des connaissances dans le domaine du patrimoine a fait l'objet de travaux ayant abouti au CIDOC CRM pour la gestion et la valorisation du patrimoine des musées (Szabados & Letricot, 2012; Le Boeuf *et al.*, 2015). Dans un cadre historique, l'apport du CIDOC CRM est central car c'est un modèle événementiel rendant compte des changements d'état. Il est donc pertinent de réutiliser le CIDOC CRM pour modéliser des ruptures et des continuités dans l'évolution historique des ports. Ces recherches sont développées dans le cadre du groupe de recherche PAM 3D Lab (où le Centre F. Viète collabore avec le Lab-STICC). Elles s'insèrent dans un projet de publication d'un corpus numérique sur la plateforme symogih.org [3], ainsi qu'un projet récent de consortium "données pour l'histoire numérique" en collaboration avec le LARHRA (UMR 5190) pour la création d'une extension "histoire" au CIDOC CRM. Dans ce cadre, la figure 1 montre un exemple simple dans le cas de l'activité humaine de "construction". Il donne le résultat d'une requête SPARQL ayant pour objet de fournir les propriétés pertinentes des quais (artefact/indicateur) et les technologies (savoirs de construction associées) pour produire une périodisation de ces ports. Elle caractérise l'évolution des quais (profondeur > 6 m) dans quatre ports : Brest en France, Rosario, Mar del Plata et Punta Alta en Argentine.

?date	?lieu	?quai	?profondeur	?longueur	?constructeur	?technologie
1859^^xsd:gYear	pho:brest	pho:quai grand mouillage b 1865	7	170	pho:Entreprise GrandHomme	pho:technologie lie au bloc de béton
1865^^xsd:gYear	pho:brest	pho:quai grand mouillage a 1873	7	165	pho:Entreprise GrandHomme	pho:technologie lie au bloc de béton
1902^^xsd:gYear	pho:rosario	pho:quai cabotage rosario	7	360	pho:entreprise Hersent	pho:technologie liée au caisson métallique
1902^^xsd:gYear	pho:rosario	pho:quai importation 1906	7	1075	pho:entreprise Hersent	pho:technologie liée au caisson métallique
1906^^xsd:gYear	pho:rosario	pho:quai importation 1912 b	7	1375	pho:entreprise Hersent	pho:technologie liée au caisson métallique
1906^^xsd:gYear	pho:rosario	pho:quai importation 1912 a	7	950	pho:entreprise Hersent	pho:technologie lie au bois
1909^^xsd:gYear	pho:mar del plata	pho:quai cabotage mdp	6.20	360	pho:SNTP	pho:technologie lie au bloc de béton
1909^^xsd:gYear	pho:mar del plata	pho:quai ultramar mdp	9	620	pho:SNTP	pho:technologie lie au bloc de béton
1912^^xsd:gYear	pho:Punta Alta	pho:quai Arroyo Pajera	9.14	300	pho:Regie generale des chemins de fer	pho:technologie liée au caisson de béton
1912^^xsd:gYear	pho:Punta Alta	pho:quai Arroyo Pajera	9.14	300	pho:entreprise Hersent	pho:technologie liée au caisson de béton

FIGURE 1 – Evolution des Quais

L'historien peut y observer des périodes de stabilité et de rupture à partir de la longueur et de la profondeur de ces quais. Pour les ports de Brest et Mar Del Palta, il peut aussi constater que la technologie utilisée perdure sur une période de 50 ans. De même, on pourrait aussi retrouver les autres savoirs mis en oeuvre et les acteurs (ingénieurs, entreprises, etc.) impliqués dans ces évolutions.

Les principales contributions de notre travail de recherche sont : i) Un modèle en SHS de périodisation des ports (HST-PORT) ; ii) Une ontologie de référence, appelée Port History Ontology (PHO) dont les classes, les propriétés et la structure sont issues du modèle HST-PORT et des sources historiques disponibles. Cette ontologie réutilise l'ontologie CIDOC CRM et DOLCE. Comme CIDOC CRM ne permet la représentation du temps (instant et intervalle) que par des chaînes de caractères, nous avons réutilisé l'ontologies "OWL-Time" pour définir les périodes de stabilité (intervalles de temps). Afin d'assurer la géolocaliation des artefacts, l'ontologie WGS84-pos nous a semblé pertinente. La suite de l'article commence par présenter le processus de construction de l'ontologie PHO, pour ensuite expliciter l'alignement des classes principales de PHO (Acteurs, Savoirs, Artefacts et Activités) avec celles des ontologies CIDOC CRM, DOLCE, OWL-TIME et WGS84-pos. Puis, l'exemple du port de Rosario est utilisé pour montrer la pertinence du modèle évènementiel (CIDOC CRM) par rapport à notre problématique. L'article se termine par une conclusion et des perspectives

3. `http://symogih.org/`

2 Conception de l'ontologie PHO

D'un point de vue SHS, la conception de l'ontologie PHO est fondée sur l'utilisation de notre modèle d'évolution des ports HST-PORT. Ce dernier présente deux aspects : i) l'activité humaine (par exemple la construction de quais) à laquelle on associe trois entités : celles des artefacts, des acteurs et des savoirs ; ii) une évolution temporelle en cinq phases : émergence d'un besoin, choix d'une solution technologique, construction, usage, disparition/obsolescence. De manière générale, la figure 2 explicite le processus de conception de l'ontologie PHO. L'historien doit avant tout retrouver et analyser les sources historiques pertinentes pour les quatre ports cités précédemment.

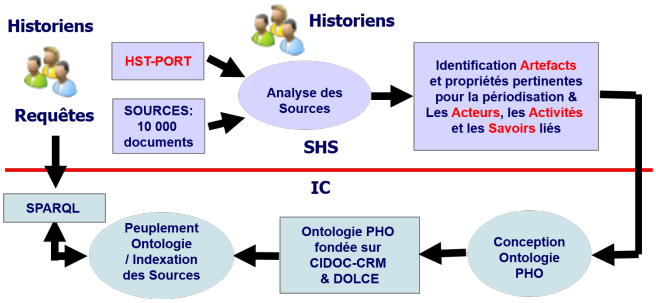


FIGURE 2 – Processus de conception

Pour chacun de ces ports, il doit identifer dans les sources les cinq phases d'évolution, trouver les artefacts et leurs propriétés pertinentes qui forment des traces tangibles des évolutions du port à différentes échelles de temps et d'espace. Pour chacun de ces artefacts, il associe les savoirs, les acteurs et les activités correspondantes. C'est donc à partir des sources historiques et du modèle HST-PORT qu'il est possible de sélectionner les différentes classes d'artefacts pertinentes et celles des activités, acteurs et savoirs associés pour concevoir l'ontologie PHO.

D'une point de vue ingénierie des connaissances, notre modéle SHS est cohérent avec l'analyse théorique de G. Kassel (Kassel, 2009). Il associe à des artefacts (de même nature que les nôtres) : des actions, des compétences et des agents qui correspondent à nos activités, savoirs et acteurs. Dans ce cadre, nous pouvons considérer les activités humaines en tant que perdurants impliquant trois classes d'endurants (artefact, acteur, savoir). Par ailleurs, Kassel a montré qu'il était plus pertinent de spécialiser une ontologie formelle comme DOLCE plutôt d'autres ontologies (Opencyc, SUMO, etc.) pour définir son ontologie formelle des artefacts. Nous reprenons donc cette approche en spécialisant de la même manière l'ontologie DOLCE. Nous allons maintenant préciser l'alignement des classes principales de l'ontologie PHO (Acteurs, Savoirs, Artefacts et Activités) avec des classes des ontologies CIDOC CRM, DOLCE, OWL-TIME et WGS84-pos.

2.1 Alignement avec DOLCE

Kassel a montré que les activités humaines sont des perdurants et que les trois classes artefacts, acteurs et savoirs sont des endurants. Les endurants sont des entités qui persistent dans le temps. Dans les endurants on distingue les "physical objects" et les "Non physical object".

Les premiers sont repérables spatialement recouvrent bien nos artefacts. La classe des "Non physical object" recouvre le domaine des entités sociales et nous pouvons y rattacher les classes "acteurs" et "savoirs" (les procédures, les savoirs scientifiques ou technologiques mis en oeuvre dans un port). Les perdurants sont des entités qui se déroulent dans le temps et les endurants participent à un perdurant durant un Time Interval (Kassel, 2009). On reconnaît donc comme perdurant notre concept d'activité. L'arborescence à gauche de la figure 3 montre l'alignement des classes "Acteurs", "Savoirs", "Artefacts", et "Activités" par rapport aux deux sous-classes d'endurant ("physical objects"et "Non physical objects") et aux perdurants. On peut aussi y voir que la classe "Artefacts" est une sous-classe de "WGS84-pos :SpatialThing" afin de répérer spatialement les actefacts. Les principales classes de PHO (Acteurs, Savoirs, Artefacts et Activités) ne sont pas uniquement des spécialisations des classes de DOLCE, mais aussi des spécialisations de classes de CIDOC CRM comme nous allons le préciser maintenant (héritage multiple).

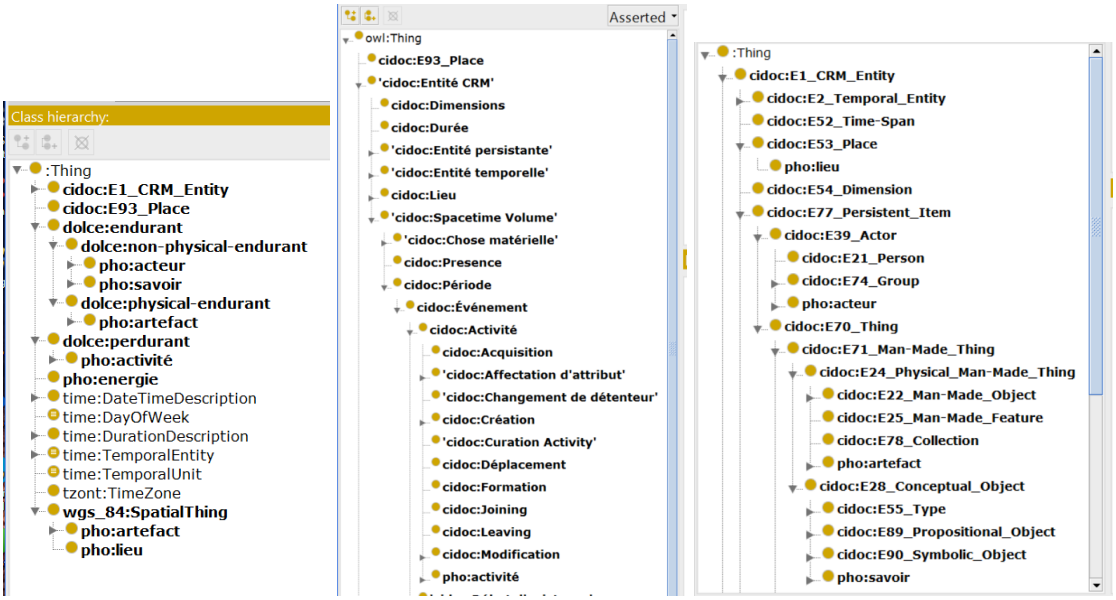


FIGURE 3 – PHO : Alignement avec CIDOC CRM et DOLCE

2.2 Alignement avec CIDOC CRM

L'arborescence centrale et de droite dans la figure 3 montre l'alignement de nos classes avec celles de CIDOC CRM. Les classes "Artefacts", "Acteurs" et "Savoirs" sont respectivement des sous classes de "E24 Physical Man Made Thing", "E39 Actor", "E28 Conceptual object". La classe "Activités" est une sous classe de "E7 Activity" - arborescence centrale. De manière plus précise :

. **Artefacts** : ce sont des productions humaines (Pomian, 2014), construites par la main de l'homme dans une intention quelconque. La classe "E24" se rapprochent fortement de la définition d'artefact que donne Pomian. La classe "Artefacts" est une spécialisation des classes "E24" du CIDOC CRM et de celle de "physical objects" de DOLCE.

. **Acteurs** : ce sont des êtres humains ou un groupe d'êtres humains. Un acteur produira

une activité via l'utilisation d'un artefact. Pour les ports, les acteurs sont très nombreux : ingénieurs, chefs de travaux, ouvriers, grutiers, les conducteurs d'engins, les manutentionnaires... Ils seront tous engagés dans une activité et la réaliseront grâce à leur savoir et à leur savoir-faire. La notion d'acteur se retrouve également dans le CIDOC CRM avec la classe "E39 Actor". La classe "Acteurs" est une spécialisation des classes "E39-Actor" du CIDOC CRM et de celle de "Non physical objects" de DOLCE.

. **Savoirs** : il peut s'agir de théories, de concepts, mais aussi de procédures comme du savoir-faire technologique. Le savoir est impliquée dans l'activité de l'acteur. L'usage d'une technologie particulière dans l'activité de construction d'un artefact peut être considéré comme un bon indicateur de la périodisation. Le CIDOC-CRM regroupe également les connaissances au sens large dans la classe "E28 Conceptual object". La classe "Savoirs" est une spécialisation des classes "E28 Conceptual object" du CIDOC-CRM et de celle de de "Non physical objects" de DOLCE.

. **Activités** : elles représentent des actions réalisées par un acteur, impliquant l'artefact, suivant des procédures ou utilisant un savoir ou un savoir-faire. Le cycle de vie d'un artefact fait appel à de nombreuses activités humaines comme la conception de l'artefact, la construction, la réparation, l'utilisation, la destruction. Une activité se déroule sur une durée ; on peut définir son début et sa fin. Ce sont les relations entre l'artefact, l'acteur et les connaissances qu'il possède qui permettront de réaliser l'activité. Le CIDOC CRM emploie aussi le concept d'activité "E7 Activity" qui spécialise la classe "Evènement" du CIDOC CRM. La classe "Activités" est donc une spécialisation des classes "E7-Activity" du CIDOC-CRM et de celle de "Perdurant" de DOLCE.

3 Evènements, activités et evolution des ports

L'ontologie CIDOC CRM est un modèle événementiel rendant compte des changements d'états, donc des évolutions des ports. Nous allons maintenant prendre l'exemple du port de Rosario pour montrer comment notre ontologie de référence PHO extension de CIDOC CRM et OWL-TIME, nous permet de rendre compte des évolutions des ports.

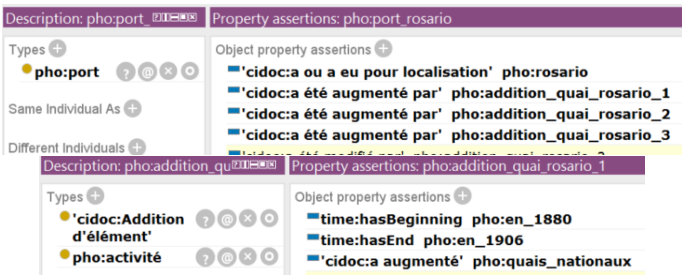


FIGURE 4 – Evolution port de Rosario, les Quais Nationaux

L'exemple est le suivant : l'Argentine a fait construire à Rosario, en 1880 des quais appelés "quais nationaux" qui seront démolis pour être remplacés en 1906 par des quais modernes appelés "quai d'importation". Puis en 1912, les quais de cabotage sont ouverts au commerce. Dans la figure 4, l'évolution d'un artefact se modélise par des activités (qui sont des évènements et des

perdurants) bornées dans le temps - un intervalle avec un instant de début et de fin. La propriété "cidoc :P110i was augmented by" permet ici de définir l'ajout de trois nouveaux artefacts (trois additions de quais : addition1, addition2, addition3) en 1880, 1906 et 1912. l'instance d'activité "pho :addition1 quai rosario" représente l'addition dans le port de Rosario d'un premier artefact : les quais nationaux. Cette activité commence "en 1880" et se termine "en 1906" qui sont deux instants de type XSD :gYear.

4 Conclusion et perspectives

L'ontologie PHO est le résultat d'une recherche pluridisciplinaire portant sur la modélisation des connaissances à l'articulation SHS/STICC. En termes de procédures scientifiques, ont été élaborés : i) en SHS : le modèle d'évolution spatio-temporel HST-PORT ; ii) en ingénierie des connaissances : l'ontologie PHO à partir de l'ontologie de référence CIDOC CRM en tant que traduction du modèle HST-PORT, PHO pouvant être considérée comme un premier élement d'une extension de CIDOC-CRM dans le domaine de l'histoire des ports.

Au delà de la thématique en histoire des ports, ce projet de recherche a aussi pour objectifs d'élaborer et valider des méthodologies de référence en humanités numériques. Tout d'abord, nous constatons une très bonne correspondance entre nos concepts avec ceux développés par Kassel en lien avec DOLCE. Par ailleurs, si ce travail montre aussi un recouvrement avec CIDOC CRM (artefact/acteur), plusieurs points cruciaux constituent des verrous à travailler de manière approndie : la question de la modélisation des savoirs et l'activité, d'une part, et, d'autre part, le fait que CIDOC CRM permet certes de décrire des événements mais présente une lacune importante pour les historiens puisqu'il ne permet pas de décrire des états d'un endurant. Ces points sont travaillés désormais dans un cadre plus large d'un projet collaboratif de création d'une extension de CIDOC CRM initié par le LARHRA (UMR 5190) à Lyon (par et pour les historiens) et par la création d'un projet de publication de corpus numérique sur l'histoire des ports sur la plateforme symogih.org.

Références

BRUNEAU O., GARLATTI S., GUEDJ M., LAUBÉ S. & LIEBER J. (2015). Semantichpst : applying semantic web principles and technologies to the history and philosophy of science and technology. In *SW4SH 2015 : First International Workshop on Semantic Web for Scientific Heritage / in conjunction with ESWC 2015 Satellite Events of the 12th International Conference on the Semantic Web*, volume 9341 of *LNCS (Lecture Notes in Computer Science)*, p. 416–427, Portoroz, Slovenia : Springer.

HUGHES T. P. *et al.* (1987). The evolution of large technological systems. *The social construction of technological systems : New directions in the sociology and history of technology*, p. 51–82.

KASSEL G. (2009). Vers une ontologie formelle des artéfacts. In *Vingtièmes Journées Francophones en ingénierie des connaissances*.

LE BOEUF P., DOERR M., ORE C. E. & STEAD S. (2015). *Definition of the CIDOC Conceptual Reference Model Documentation Standards Group V6*. Rapport interne, CIDOC International Committee.

POMIAN K. (2014). De l'exception humaine. *Le Débat*, **3**(180), 31–34.

SZABADOS A.-V. & LETRICOT R. (2012). L'ontologie CIDOC CRM appliquée aux objets du patrimoine antique. In *Troisièmes Journées d'Informatique et Archéologie de Paris - JIAP 2012*, Paris, France.

OntoCoins : données ouvertes liées pour la numismatique, patrimoine culturel

Cédric Lopez [1], Marie-Laure Le Brazidec [2], Jean-Albert Chevillon [3], Francis Couturas [4], Dominique Hollard [5], Aurélien Pierre [6]

[1] UMR8546 CNRS/ENS Archéologie et Philologie d'Orient et d'Occident – AOROC, Paris
[2] Musée Saint-Raymond, Toulouse
[3] Musée archéologique du Pègue, Le Pègue
[4]Musée d'art et d'archéologie de Périgueux, Périgueux
[5]Bibliothèque nationale de France, Paris
[6]Musée Fenaille, Rodez

Résumé : Le patrimoine culturel est en partie conservé dans des collections publiques et privées difficilement accessibles au grand public ainsi qu'aux chercheurs, historiens de l'art, économistes, ou archéologues, qui trouvent là un frein majeur à leurs recherches. Nous sommes aujourd'hui en mesure de partager et de valoriser ce patrimoine par le biais du Web sémantique et des données ouvertes liées sous l'impulsion du Ministère de la Culture. Dans cet article, nous développons un vocabulaire qui a pour vocation d'ouvrir les données « numismatiques », c'est-à-dire des données relatives aux monnaies anciennes, véritables œuvres d'art réalisés par des artistes graveurs, témoins de l'évolution des cultures depuis plus de deux millénaires, dont seule une petite partie (dans le meilleur des cas) est exposée au public à travers les vitrines des musées. La conception et le développement de notre vocabulaire baptisé OntoCoins sont traités dans cet article. L'application Wikimoneda a été développée pour montrer la pertinence du vocabulaire, et est d'ores et déjà alimentée (et consultée) par cinq des plus importants conservateurs de collections publiques collaborant dans ce travail.

Mots-clés : Modèles de connaissances, Conception et réutilisation d'ontologies, Web sémantique, Web des données, Approches interdisciplinaires.

1 Introduction

Le Ministère de la Culture montre des signes encourageants concernant l'ouverture et la mise à disposition des données culturelles. L'article 11 de la loi du 17 juillet 1978 (loi CADA) inclus une disposition appelée « exception culturelle » qui permet à chaque établissement, organisme et service culturel de fixer ses propres conditions de réutilisation des données[1]. Très récemment, la loi du 28 septembre 2016 a créé l'obligation pour les organismes publics de communiquer gratuitement sur internet leurs bases de données, sous réserve d'anonymisation et de protection du secret industriel et commercial, qui pourront ainsi être exploitées et réutilisées facilement par un particulier.
Une partie du patrimoine historique et culturel concerne la numismatique, une science de l'archéologie qui étudie les monnaies dont l'apparition remonte au VIIe siècle avant notre ère. Des centaines de milliers de monnaies sont conservées dans le monde, dans des médailliers, *i.e.* des meubles à tiroirs accueillant chacun des cases où sont logées les pièces. A chaque

[1] Article 11 : « Par dérogation au présent chapitre, les conditions dans lesquelles les informations peuvent être réutilisées sont fixées, le cas échéant, par les administrations mentionnées aux a et b du présent article lorsqu'elles figurent dans des documents produits ou reçus par : a) Des établissements et institutions d'enseignement et de recherche ; b) Des établissements, organismes ou services culturels ».

pièce est associée, par le biais d'une étiquette, un certain nombre d'information (références, masse, diamètre, métal, autorité émettrice, *etc.*). Ces médailliers ne contiennent ni des œuvres protégées par un droit de propriété littéraire et artistique ou industrielle ni des données personnelles ; ce ne sont pas des données sensibles.
Parce que l'accès aux médailliers demande une logistique coûteuse pour les organismes publics et que les monnaies sont réparties dans différents lieux, les chercheurs en Archéologie et en Histoire de l'Art, entre autres, sont freinés dans leurs travaux dont la phase d'acquisition des données ne peut être aisément conduite.
Nous proposons dans cet article le vocabulaire OntoCoins qui a pour ambition de représenter les données relatives à la numismatique dans le format des données liées. Ce vocabulaire, conçu en collaboration avec le Musée Saint-Raymond (Toulouse), le Musée d'art et d'archéologie de Périgueux, la Bibliothèque nationale de France, le Musée archéologique du Pègue et le Musée Fenaille (Rodez) doit permettre de représenter puis d'interroger des connaissances relatives aux monnaies (leur poids, leur description, *etc.*), aux trésors, aux découvertes archéologiques, aux études de coins monétaires (i.e. les matrices ayant servies à fabriquer les monnaies), aux prototypes, aux styles de gravures, *etc.*
Ce vocabulaire a un double objectif : d'une part, il s'agit de libérer les informations préservées dans les médailliers depuis parfois plusieurs siècles pour les rendre accessible au public ; d'autre part, il s'agit de capitaliser les résultats de recherche menés ces dernières décennies, en particulier ceux obtenus à partir de la « charactéroscopie », une méthode visant à reconnaître les marques distinctives de chaque coin monétaire afin d'identifier des monnaies frappées dans un même atelier monétaire (Colbert de Beaulieu, 1973).
En s'appuyant sur le vocabulaire OntoCoins, nous avons développé l'application Wikimoneda qui permet de consulter les monnaies gauloises (dans un premier temps) conservées dans les musées partenaires et de faire apparaître des informations qui n'apparaissaient pas au premier abord. Plus techniquement, une telle application serait en mesure de répondre, par exemple, à la requête suivante : « Quels sont les coins monétaires utilisés pour la fabrication des monnaies des trésors de Dunes et de Saint-Etienne-des-Landes ? Combien de monnaies connait-on qui soient issues de ces coins monétaire ? Quelles sont leurs provenances ? ».
Dans la suite de l'article, nous présentons les travaux antérieurs (section 2), puis nous suivons la méthodologie de Grüninger et Fox (1995) pour construire le vocabulaire OntoCoins (section 3). Nous montrons la pertinence du vocabulaire développé par son utilisation dans l'application Wikimoneda (section 4).

2 Travaux antérieurs

Dans ses réflexions sur les bases de données archéologiques, (Feugère, 2015) met en avant que l'archéologie « accuse un certain retard sur d'autres sciences humaines en matière de bases de données », mais la motivation est forte. Il existe aujourd'hui un réel engouement pour l'ouverture des données archéologiques, et plus précisément numismatiques. Ripollès et Gozalbes (2014) dressent un état de la question des collections publiques de monnaies antiques en ligne et en recensent 26, celles-ci n'étant pas au format des données liées. D'autres projets s'appuient sur les recommandations de CIDOC-ICOM, en particulier sur LIDO (Cuburn et al., 2010), un schéma XML permettant la description d'une large gamme d'informations sur les objets conservés dans les musées (géologie, art, botanique, archéologie, *etc.*).
Il semble qu'un seul vocabulaire ait émergé pour représenter les connaissances numismatiques, à l'initiative de l'American Numismatic Society en 2010. L'objectif du vocabulaire Nomisma[2] est d'ouvrir les données numismatiques sous forme de catalogue (Gruber et al., 2012). Le vocabulaire est constitué de 30 classes (par exemple *Collection*, *Mint*, *Material, Controlmark*) et de 52 propriétés (par exemple *hasMaxDiameter,*

[2] http://nomisma.org/

nmo:hasLegend, hasAuthority, *hasDate*, *hasFindSpot*). Son utilisation par des acteurs tels qu'Online Coins of the Roman Empire, Antike Fundmuenzen Europa, ou Portable Antiquities Scheme, démontre l'intérêt du vocabulaire pour représenter des connaissances numismatiques. Cependant, Nomisma ne couvre pas tous les concepts numismatiques. En particulier, ceux relatifs aux études de coins monétaires et aux spécificités des monnayages, sont absents puisqu'il se limite à représenter les descriptions de monnaies.
Nomisma présente l'avantage de laisser une totale liberté à l'utilisateur quant à la modélisation de ses connaissances : *rdf:domain* et *rdf:range* ne sont pas définis. Ainsi, si deux utilisateurs font des choix d'utilisation différents, les résultats de requêtes ne seront, dans certains cas, pas pertinents. L'interopérabilité entre les bases de connaissance n'est donc pas assurée. Ainsi, conscients des problématiques d'interopérabilité que leur choix engendre (Tolle *et al.*, 2016), la priorité a délibérément été donnée à la liberté de l'utilisateur. Dans notre cas, nous faisons le choix contraire. L'interopérabilité entre bases de connaissances doit selon nous être prioritaire, quitte à poser des contraintes d'utilisation plus fortes. Cette opposition nous a conduits à développer le vocabulaire OntoCoins. Nous conservons de Nomisma la plupart des noms de propriétés utiles à la représentation de données de type « catalogue » en définissant leurs domaines et co-domaines.
Pour développer le vocabulaire OntoCoins, nous avons suivi la méthodologie proposée par Grüninger et Fox (1995). Celle-ci se compose de trois phases : 1) l'élaboration de scénarios, c'est-à-dire la description de cas d'utilisation réels, 2) l'écriture des questions de compétences qui correspondent aux questions de l'utilisateur auxquelles le système automatique doit être en mesure de répondre, 3) le choix du vocabulaire avec lequel seront exprimées les requêtes du numismate.

3 Le vocabulaire OntoCoins

Dans cette section, nous exposons notre travail de conception et de développement du vocabulaire selon la méthodologie de Grüninger et Fox (1995).

3.1 Scénarios

Les scénarios donnés ici en exemple sont issus des besoins communs des chercheurs numismates et des conservateurs du patrimoine impliqués dans ce travail. Ces scénarios motivent et délimitent la conception et la publication du vocabulaire OntoCoins et permettent d'identifier des applications associées. La liste de ces scénarios n'est pas exhaustive.

- Scénario 1 : L'utilisateur recherche une monnaie. Le processus de recherche nécessite une connaissance d'informations telles que : entité émettrice, période de fabrication, représentation et description complète de l'empreinte, poids moyen, *etc.* Le vocabulaire doit donc permettre de représenter une fiche descriptive pour chaque monnaie.

- Scénario 2 : L'utilisateur souhaite connaître les lieux de découverte des monnaies. Des cartes géographiques permettent de visualiser les provenances. Le vocabulaire doit donc permettre de représenter les lieux de provenances.

- Scénario 3 : L'utilisateur souhaite connaître le nombre de coins monétaires (matrices) recensés pour un type donné. Le vocabulaire doit donc permettre de représenter les coins monétaires et des informations relatives.

- Scénario 4 : L'utilisateur souhaite connaître le poids moyen des monnaies issues d'un (ensemble de) coin(s) monétaire(s). Le vocabulaire doit donc permettre de représenter des valeurs quantitatives pour mesurer une masse ou une dimension.

- Scénario 5 : L'utilisateur souhaite connaître les liaisons de coins entre des trésors monétaires. (5bis) L'utilisateur souhaite également connaître l'ensemble des monnaies frappées au sein d'un atelier monétaire. Le vocabulaire doit donc permettre de représenter les liaisons de coins.

- Scénario 6 : L'utilisateur souhaite connaître la bibliographie qui évoque une monnaie, un trésor, … Le vocabulaire doit donc permettre l'association de monnaies, trésor, etc, à la bibliographie.

- Scénario 7 : L'utilisateur souhaite connaître le(s) prototype(s) d'une monnaie. Le vocabulaire doit donc permettre de représenter les prototypes, qu'ils soient locaux ou non.

3.2 Questions de compétences

Afin de déterminer les spécifications de l'ontologie, nous dressons une liste de questions de compétences tel que recommandé par (Grüninger & Fox, 1995). Une liste non exhaustive des questions de compétence issues des scénarios précédents est présentée :

- Q1 : Quelles monnaies montrant une hache sur une face ont été frappées entre les années 120 et 52 ? (Scénario 1)
- Q2 : Où sont localisées les découvertes de monnaies montrant une hache ? (scénario 3)
- Q3 : Où sont localisés les trésors qui contiennent au moins une monnaie représentant une hache ? (Scénarios 1 et 2)
- Q6 : Combien de monnaies parvenues jusqu'à nous ont été produites par le coin de revers n°164 ? et au sein d'un même atelier ? (Scénario 5)
- Q7 : Quel est le poids des monnaies frappées par les coins de revers n°164, 165 et 166 ? (Scénario 4)
- Q8 : Quelles sont les publications qui mentionnent le trésor de Pinsaguel ? (Scénario 6)
- Q9 : Quels sont les prototypes d'une gravure ? (Scénario 7)

3.3 Modélisation

Dans cette section nous présentons et discutons les principales entités et relations du vocabulaire OntoCoins. Elles sont issues de la terminologie extraite des questions de compétences, lorsque celles-ci n'étaient pas représentées dans d'autres vocabulaires tels que

Nomisma. En ce qui concerne le vocabulaire défini dans Nomisma, nous en faisons usage dès lors qu'il nous parait pertinent pour notre modèle. De cette façon, Nomisma et OntoCoins sont liés et couvrent ensemble une représentation plus large des connaissances numismatiques.
La recherche des vocabulaires déjà existants (vocabulaire pour exprimer des informations spatiales ou des données relatives à la bibliographie, par exemple) s'est principalement effectuée via Linked Open Vocabularies (LOV) : http://lov.okfn.org/dataset/lov/. Les primitives centrales du vocabulaire OntoCoins sont présentées en Figure 1 –.
Dans la suite, le préfixe *oc:* désigne des ressources de OntoCoins (espace de nommage : http://purl.org/ontocoins/v1), et le préfixe *nmo:* désigne des ressources de Nomisma. L'espace de nommage et la publication respectent les principes des données liées sur le Web et notamment la déréférenciation et la négociation de contenu par HTTP. Nous présentons dans la suite quelques classes de l'ontologie OntoCoins et nous justifions les choix réalisés. Un travail similaire sur la modélisation des propriétés a été mené. Le cœur du vocabulaire est consultable en Figure 1.

Exemples de classes

- **oc:Coin :** La classe « Coin » permet de représenter des monnaies. Cette classe n'est pas définie dans Nomisma qui propose l'utilisation de la classe NumismaticObject définie ainsi : « The physical objects that are of interest in the numismatic domain. » Il nous semble plus pertinent de définir une classe « Coin », avec la possibilité de définir d'autres classes destinées à représenter individuellement chaque type d'artefact numismatique (coins monétaires, poids monétaires, par exemple). Notre classe wm:Coin doit donc être considérée comme une sous-classe de la classe nmo:NumismaticObject.
- **oc:Hoard :** La classe "Hoard" permet de représenter des trésors. Cette classe permet de représenter un trésor comme une ressource à part entière. Cette classe n'est pas définie dans Nomisma où l'on représente le fait qu'une monnaie « dcterms:isPartOf » (est une partie de) un trésor, mais le trésor ne pouvant être rattaché à une classe lui correspondant, des ambiguïtés peuvent apparaître, par exemple, une monnaie peut également être *une partie d'*un bijou ». L'objet n'est donc pas toujours un trésor, d'où la nécessité de la classe wm:Hoard que nous définissons dans OntoCoins.
- **oc:Obverse :** La classe « Obverse » représente des droits (côté « pile » de la monnaie). Cette classe n'est pas définie dans Nomisma. Il nous semble pertinent de définir une telle classe pour distinguer sans ambiguïté possible les droits des revers (côté « face » de la monnaie). Cette distinction peut néanmoins être faite dans Nomisma en utilisant les propriétés nmo:hasObverse et nmo:hasReverse, mais l'utilisation de ces propriétés dont les domaines et co-domaines ne sont pas définis permettent une utilisation plus large, par exemple pour une médaille, ou tout autre objet à deux faces, ce que nous souhaitons éviter.
- **oc:Reverse :** La classe "Reverse" représente des revers. Commentaires similaires à ceux de la classe wm:Obverse.

- **oc :ArcheologicalSite :** La classe « ArcheologicalSite » représente les sites archéologiques. Voir les commentaires de la classe wm :StratigraphicUnit.
- **oc :StratigraphicUnit :** La classe " StratigraphicUnit " représente les unités stratigraphiques (terme archéologique). Cette classe n'est pas définie dans Nomisma qui a pour seul moyen de localisation la propriété nmo:hasFindspot (« Describes the location of the discovery of an object, whether by accident or in archaeological excavation. »). Dans OntoCoins, nous avons représenté les unités stratigraphiques, les sites archéologiques, et les villes.

3.4 Choix du langage et évaluation

OntoCoins a été édité avec le logiciel Protégé[3]. Le vocabulaire et la description de ses ressources et propriétés sont publiées selon les principes des données liées sur le Web et le schéma est identifié par l'URI http://purl.org/ontocoins/v1. Le vocabulaire déjà défini par ailleurs a été adopté le cas échéant (FOAF, EXIF, BIBO, DUBLIN CORE, *etc.*).
OntoCoins utilise les mêmes primitives que Nomisma : owl:Ontology, owl:Class, owl:DatatypeProperty, rdfs:subClassOf, rdfs:subPropertyOf, rdfs:comment, rdf:type. De cette façon, les annotations effectuées avec OntoCoins et Nomisma sont dans le même fragment d'expressivité et peuvent être interprétées par un raisonneur RDF(S) qui sait traiter les éléments mentionnés ci-dessus.
Nous avons montré, par le biais de quelques exemples de requêtes SPARQL, que toutes nos questions de compétences trouvent une réponse. Quelques exemples de requêtes sont donnés dans la suite, pour lesquelles on notera le préfixe *oc: <http://purl.org/ontocoins/v1#>.*

- Pour la question de compétence Q1 «Quelles monnaies avec une hache ont été frappées entre les années -120 et -52 ?» issue du scénario 1, des exemples de questions concrètes sa formulation en SPARQL est :

```
SELECT DISTINCT ?coin
WHERE {
    ?coin  oc:hasReverse ?reverse .
    ?coin oc:terminusPostquem "-120" .
    ?coin oc:terminusAntequem "-52" .
    ?reverse dc:description ?description .
FILTER regex( ?description, "hache" )
}
```

[3] http://protege.stanford.edu/

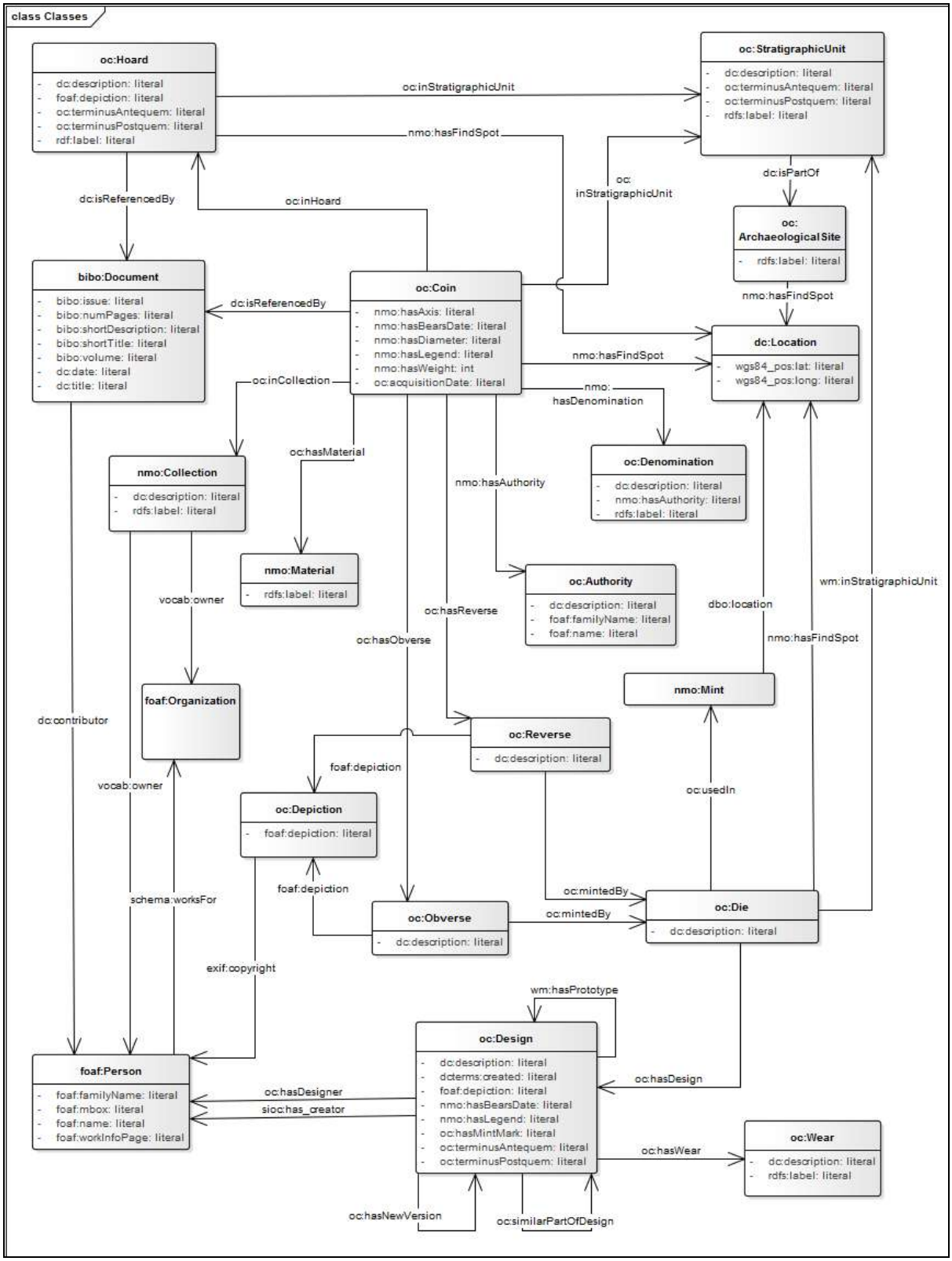


FIGURE 1 – Aperçu du vocabulaire OntoCoins (non exhaustif).

- Pour la question de compétence Q2 «Où sont localisées les découvertes de monnaies montrant une hache ?» issue des scénarios 1 et 2, des exemples de questions concrètes sa formulation en SPARQL est :

```
SELECT DISTINCT ?geoloc
WHERE {
    ?coin dbo[4]:location ?geoloc .
    ?coin  oc:hasReverse ?reverse .
    ?reverse dc:description ?description .
     FILTER regex( ?description, "hache" )
}
```

- Pour la question de compétence Q3 «Où sont localisés les trésors qui contiennent au moins une monnaie avec une hache ? » issue des scénarios 1 et 3, des exemples de questions concrètes sa formulation en SPARQL est :

```
SELECT DISTINCT ?geoloc
WHERE {
    ?hoard dbo:location ?geoloc .
    ?coin   oc:inHoard  ?hoard .
    ?coin  oc:hasReverse ?reverse .
    ?reverse dc:description ?description .
    FILTER regex( ?description, "hache" )
}
```

- Pour la question de compétence Q4 «Combien de monnaies parvenues jusqu'à nous ont été produites par le coin de revers n°164 ?» issue du scénario 5, des exemples de questions concrètes sa formulation en SPARQL est :

```
SELECT DISTINCT ?coin (count(distinct ? coin) as ?count)
WHERE {
    ?coin  oc:hasReverse ?reverse .
    ?reverse oc:mintedBy  <http://www.wikimoneda.com/kb/Die/164>.
}
```

4 Wikimoneda.com : un exemple d'application

Basé sur OntoCoins, nous avons développé une application Web nommée Wikimoneda (Figure 2). L'application, actuellement en version bêta, est accessible via Internet : http://www.sw.wikimoneda.com/. Cette application a pour ambition de réaliser les scénarios décrits en section 3.1.

[4] http://dbpedia.org/ontology/

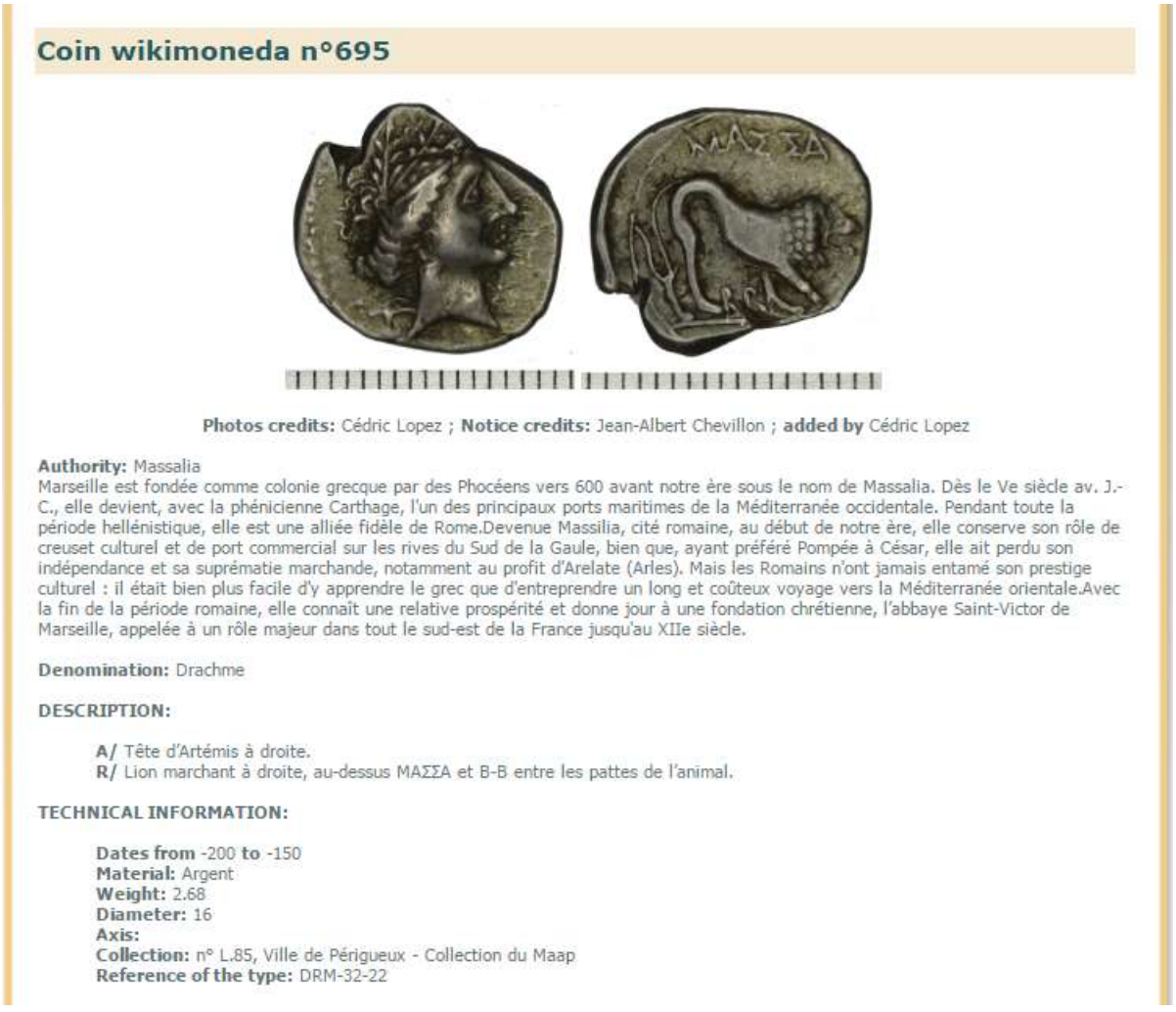
FIGURE 2 – Aperçu de la page d'accueil de l'application Wikimoneda.

Une fois connecté, l'utilisateur accède à 5 onglets :

- *Home*, page d'accueil. L'utilisateur est informé des derniers évènements liés à Wikimoneda.
- *Add*, page pour ajouter des données. L'utilisateur peut facilement structurer ses données au format RDF en remplissant de simples formulaires. Les données sont immédiatement insérées dans la base et peuvent dès lors apparaître dans les résultats de recherche.
- *Search*, page pour rechercher des données. L'utilisateur peut rechercher des données selon plusieurs critères liés aux monnaies, ou aux reconstitutions d'empreintes, par exemple : autorité émettrice, dénomination, matériau, trésor, collection, description, etc.
- *Informations*, page d'informations. L'utilisateur peut s'informer sur la démarche entreprise, peut consulter le vocabulaire OntoCoins, télécharger des données parmi 9 formats possibles (XML, JSON, Plain, Serialized PHP, Turtle, RDF/XML, Query Structure, HTML Table, TSV).

- *Login/Logout,* connexion/déconnexion. L'utilisateur peut se connecter et se déconnecter en accédant à cet onglet.

Afin de valider l'approche et l'intérêt d'un tel outil, le Musée Saint-Raymond (Toulouse), le Musée d'art et d'archéologie de Périgueux, la Bibliothèque nationale de France (Paris), le musée archéologique du Pègue et le Musée Fenaille (Rodez) ont ouvert une partie de leurs médailliers sur Wikimoneda. Au total, plus d'un millier de monnaies gauloises a été inséré dans la base, représentant une vingtaine de milliers de triplets RDF. Un exemple de fiche monétaire est visible en Figure 3.



Coin wikimoneda n°695

Photos credits: Cédric Lopez ; **Notice credits:** Jean-Albert Chevillon ; **added by** Cédric Lopez

Authority: Massalia
Marseille est fondée comme colonie grecque par des Phocéens vers 600 avant notre ère sous le nom de Massalia. Dès le Ve siècle av. J.-C., elle devient, avec la phénicienne Carthage, l'un des principaux ports maritimes de la Méditerranée occidentale. Pendant toute la période hellénistique, elle est une alliée fidèle de Rome.Devenue Massilia, cité romaine, au début de notre ère, elle conserve son rôle de creuset culturel et de port commercial sur les rives du Sud de la Gaule, bien que, ayant préféré Pompée à César, elle ait perdu son indépendance et sa suprématie marchande, notamment au profit d'Arelate (Arles). Mais les Romains n'ont jamais entamé son prestige culturel : il était bien plus facile d'y apprendre le grec que d'entreprendre un long et coûteux voyage vers la Méditerranée orientale.Avec la fin de la période romaine, elle connaît une relative prospérité et donne jour à une fondation chrétienne, l'abbaye Saint-Victor de Marseille, appelée à un rôle majeur dans tout le sud-est de la France jusqu'au XIIe siècle.

Denomination: Drachme

DESCRIPTION:

A/ Tête d'Artémis à droite.
R/ Lion marchant à droite, au-dessus MAΣΣA et B-B entre les pattes de l'animal.

TECHNICAL INFORMATION:

Dates from -200 **to** -150
Material: Argent
Weight: 2.68
Diameter: 16
Axis:
Collection: n° L.85, Ville de Périgueux - Collection du Maap
Reference of the type: DRM-32-22

FIGURE 3 – Partie d'une fiche descriptive d'une monnaie du IIe siècle avant notre ère.

Nous décrivons dans la suite un cas d'utilisation réel qui, étant donnée une monnaie M, permet de trouver toutes les monnaies frappées dans le même atelier que la monnaie M :

1. L'utilisateur recherche les monnaies gauloises montrant un sanglier au revers.
2. L'utilisateur accède à la fiche complète de la monnaie n°131.

3. La fiche contient une représentation des coins monétaires ayant frappé chaque face. L'utilisateur clique sur l'un des coins.
4. L'utilisateur accède à la fiche descriptive du coin monétaire (cf. Fig. 4) qui indique le nombre de monnaies fabriquées par ce coin. L'utilisateur a également la possibilité de générer une carte de répartition géographique. Le système renvoie ensuite tous les autres coins utilisés dans le même atelier que le coin initial.
5. Enfin, l'utilisateur accès à la liste des monnaies frappées par l'ensemble de ces coins monétaires, pour un total de 54 coins utilisés au sein d'un même atelier de fabrication. L'utilisateur peut à ce stade générer une nouvelle carte de répartition (cf. Fig. 4).

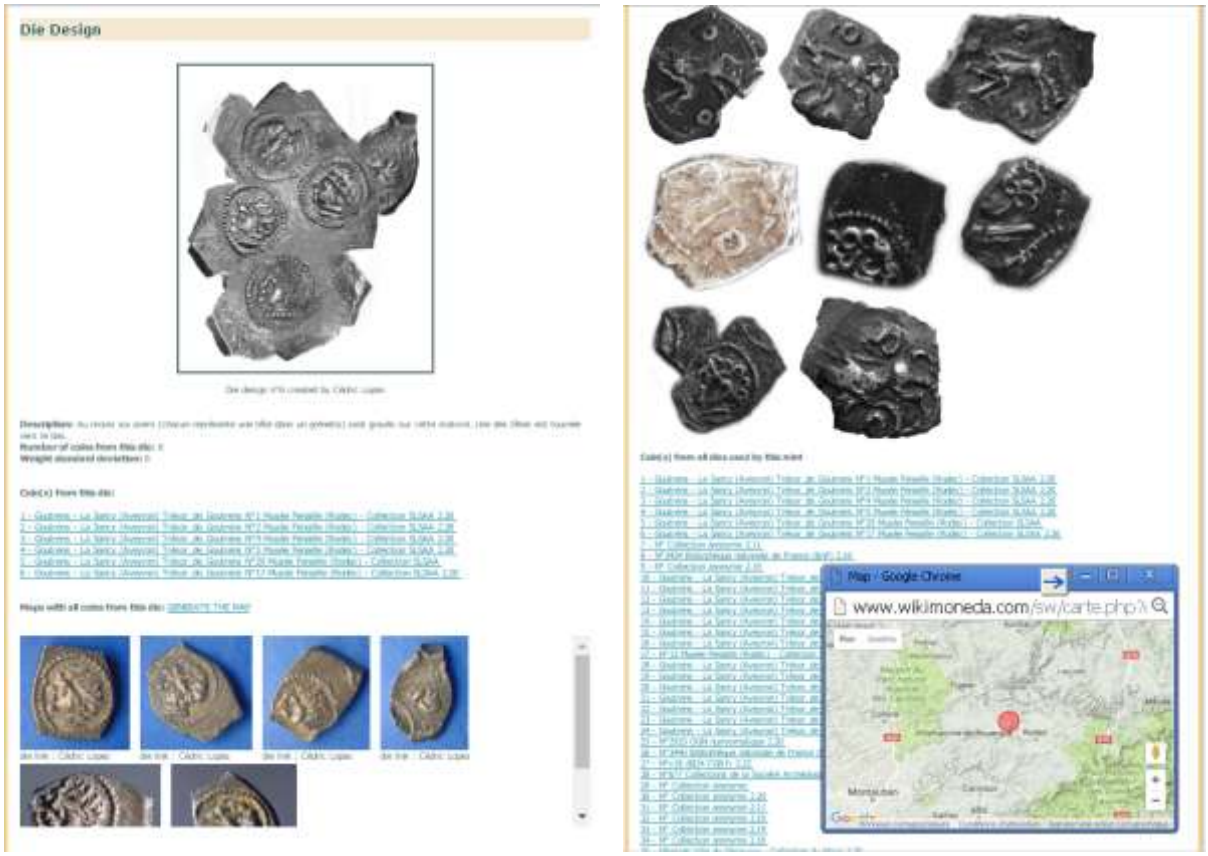


FIGURE 4 – Illustrations des scénarios ; à gauche, fiche descriptive du coin monétaire avec les informations associées ; à droite, ensemble des coins monétaires utilisés dans le même atelier que la monnaie initialement recherchée, et carte de répartition géographique de la totalité des monnaies frappées dans ledit atelier.

Ainsi, à partir d'une monnaie, le système a permis de faire émerger des liens qui n'apparaissaient pas au premier abord, témoins de la production d'un atelier monétaire du IIe siècle avant notre ère.

5 Conclusion

Fondée sur OntoCoins et Nomisma, nous avons développé et expérimenté une application concernant la numismatique. Cette application a un triple intérêt. Tout d'abord, l'application permet à des non informaticiens de transformer leurs données au format du web sémantique (RDF) via les formulaires en ligne, en toute simplicité. D'autre part, il s'agit de faciliter le partage des données avec le « grand public » dans un format interrogeable par tous et permettant de croiser des données de différentes collections, publiques et privées. Enfin, il s'agit de valoriser ces données brutes en exploitant automatiquement les liens qu'elles entretiennent pour faire émerger des connaissances qui n'apparaissent pas au premier abord (par exemple la localisation d'ateliers monétaires).

L'ensemble de ces connaissances, dorénavant accessible à tous les chercheurs en quelques clics à travers l'application Wikimoneda[5], permettra certainement un gain de temps considérable dans ces travaux de recherche.

OntoCoins est centré sur la classe « Coin » (*i.e. monnaie*). A court terme, cette classe, ainsi que la classe « Die » (*i.e. coin monétaire*) pourrait être liée à une classe plus générale « archaeological artefact ». De là, de nombreux objets archéologiques pourraient être liés. Une requête telle que « Quelles sont les monnaies découvertes dans une unité stratigraphique contenant des fragments d'amphores de type Dressel 1B ? » est envisagée dans la continuité de la démarche présentée dans cet article.

Références

COBURN E., LIGHT R., MCKENNA G., STEIN R., VITZTHUM, A. (2010) LIDO – Lightweight information describing objects.

COLBERT DE BEAULIEU, J.-B. (1973) *Traité de numismatique celtique*. Les Belles Lettres.

FEUGERE, M. (2015). Bases de données en archéologie: de la révolution informatique au changement de paradigme. *Cahiers philosophiques*, (141), p. 139-147.

GRUBER E., HEATH S., MEADOWS A., PETT D., TOLLE K., & WIGG-WOLF, D. (2012). Semantic Web Technologies Applied to Numismatic Collections. *CAA* .

GRÜNINGER, M., & FOX, M. S. (1995). *Methodology for the design and evaluation of ontologies.*

LOPEZ, C., & RICHARD RALITE, J.-C. (2014). Technique moderne de reconstitution d'empreintes monétaires: application à un type monétaire préaugustéen des Rutènes. *Etudes celtiques*(40), p. 7-20.

RIPOLLES P.-.P & GOZALBES M. (2014). La numismática de la Antigüedad online. Situación actual y perspectivas de futuro. *Actas del XV Congreso Nacional de Numismática*, p. 743-752

TOLLE K., WIGG-WOLF D., GRUBER E. (2016). An Ontology for a Numismatic Island with Bridges to Others, Acte de *Computer Applications and Quantitative Methods in Archaeology* (CAA), à paraître.

USCHOLD, M., & GRUNINGER, M. (1996). Ontologies: Principles, methods and applications. *The knowledge engineering review, 11*(2), p. 93-136.

[5] www.sw.wikimoneda.com

Ontologie modulaire pour la fabrication et l'exploitation de vêtements intelligents dédiés au sport

Samya Sagar1, Issam Rebai1, Maha Khemaja2 et Jamel Feki3

[1] Lab-STICC, IMT-Atlantique, Bretagne Loire, F-29238 Brest,
{samya.sagar, issam.rebai}@ imt-atlantique.fr
[2] Prince, Istic, Université de Sousse, Tunisie,
maha_khemaja@yahoo.fr
[3] MIRACL, Université de Sfax, Tunisie,
Jamel.feki@fsegs.rnu.tn

Résumé : Les vêtements intelligents sont des vêtements dotés de capteurs. Dans le domaine du sport, ils permettent de collecter des informations sur l'état de forme des sportifs. Dans cet article et dans le cadre des ontologies modulaires, nous présentons l'ontologie modulaire SMS conçue pour l'exploitation et la fabrication des vêtements intelligents dédiés au sport. Nous décrivons le cadre méthodologique pour la construction des différents modules. Nous donnons un aperçu des modules requis dans la phase d'exploitation et leurs relations sémantiques dans la structure de l'ontologie SMS.

Mots-clés : Ontologie modulaire, modules, réutilisation d'ontologies, vêtements intelligents de sport.

1 Introduction

Les Vêtements Intelligents (VI) sont des vêtements ordinaires augmentés par des capteurs (Rantanen & al. 2002) mesurant des grandeurs physiologiques ou actimétriques. Dans cet article, nous traitons l'usage des VI pour le sport, définis dans le cadre du projet SmartSensing[1]. Ce dernier a pour objectif de concevoir des vêtements intelligents et communicants pour le monitoring et le coaching sportifs. Plus particulièrement, notre travail s'intéresse à la modélisation sémantique de la fabrication et de l'exploitation des VI. L'intérêt de cette modélisation est de (i) monitorer les sportifs ; (ii) réutiliser un VI dans plusieurs activités sportives ; (iii) apporter de l'aide à la conception d'un VI.

L'objectif de cet article est de présenter l'ontologie modulaire SMS créée dans le cadre de ce projet. SMS répond aux différents besoins des phases de fabrication et d'exploitation d'un VI de sport. Elle est constituée de différents domaines de connaissances décrivant un tel vêtement (capteur, vêtement, sport, etc.). Chaque domaine est décrit par un module de l'ontologie. SMS réutilise l'ontologie de haut niveau DOLCE Ultra Lite (UL) ainsi que l'ontologie de capteurs SSN (*Semantic Sensor Network*) (Compton et al., 2012) recommandée par le W3C pour les projets IoT (*Internet of Things*). SSN est aussi alignée à DOLCE UL. Il s'agit ainsi de favoriser la réutilisation de l'ontologie SMS et son interopérabilité notamment dans le cadre de l'IoT. Nous esquissons, dans la section 2, le cadre méthodologique adopté pour la construction de SMS. Nous décrivons, dans la section 3, la modélisation d'un VI. La section 4 porte sur le rôle des modules identifiés de SMS. Nous présentons, dans la section 5, trois de ces modules utilisés durant la phase d'exploitation et leur structuration.

[1] http://www.smartsensing.fr/ : projet financé par Bpifrance et assuré par un consortium d'entreprises et de laboratoires de recherche.

2 Principe et méthodologie de construction de l'ontologie modulaire SMS

L'analyse des besoins du projet SmartSensing a mis en exergue la diversité des connaissances associées aux différents domaines intervenant dans le projet. Cette diversité entraîne une complexité au niveau conceptuel et organisationnel de ces connaissances. Adopter la « modularisation » comme principe de construction de l'ontologie SMS, consiste à découper les connaissances du domaine en modules, contenant différents types de connaissances (Borst, 1997). Selon (D'Aquin et al., 2007) la « modularisation » est un moyen de structurer et d'organiser des ontologies et permet de construire de larges ontologies, fondées sur la combinaison de composantes de connaissances autonomes, indépendantes et réutilisables. Dans notre travail, nous considérons que ces composantes sont elles-mêmes des ontologies, dites « modules », et que l'ontologie résultante de cette composition est une « ontologie modulaire ». Chaque module couvre un domaine d'expertise particulier pour la fabrication et l'exploitation d'un VI de sport (capteurs, vêtements, etc.). Ainsi, il est possible de réutiliser les différents modules indépendamment les uns des autres. En changeant le domaine de sport, par exemple, nous réutilisons le VI pour d'autres domaines d'application, comme le militaire, la santé, etc.

La conception modulaire poursuit deux autres objectifs : (1) favoriser leur réutilisation par d'autres ; (2) réutiliser des ontologies de référence (SSN, DOLCE Ultra Lite, etc.). La réutilisation d'ontologies est une pratique fortement recommandée dans la communauté du Web sémantique[2] (Bizer et al., 2009). Elle offre la possibilité de : (1) réduire le coût de création d'ontologies, (2) améliorer la qualité des ontologies résultantes et (3) faciliter l'interaction ultérieure entre les systèmes (Stecher et al., 2008). La réutilisation ontologique peut avoir plusieurs formes ; les ontologies peuvent être référencées, importées, prises comme point de départ pour des extensions, des révisions, etc. (Stecher et al., 2008) distinguent trois types de réutilisation d'ontologies : (1) réutilisation conservatrice, (2) réutilisation adaptative et (3) réutilisation de la bonne pratique. Pour créer des ontologies modulaires basées sur la réutilisation, le projet Neon[3] propose un ensemble de méthodologies (Suarez-Figueroa et al., 2012), définies à travers neuf scénarios. Dans notre travail, nous avons appliqué les 4 scénarios suivants :

Le scénario 1 (De la spécification à l'implémentation), présentant le processus classique de construction d'ontologies, que nous avons respecté pour la construction de chaque module.

Le scénario 2 (Réutilisation et réingénierie des ressources non-ontologiques), où l'analyse des ressources disponibles, fournies par les différents experts du projet, nous a permis de modéliser les connaissances relatives aux différents modules de l'ontologie SMS.

Le scénario 3 (Réutilisation des ressources ontologiques), a été appliqué pour une réutilisation conservatrice d'ontologies de référence. Au niveau d'abstraction le plus élevé de l'ontologie SMS, nous avons réutilisé l'ontologie DOLCE UL. La conception des modules a été fondée sur la spécialisation de cette ontologie. L'ontologie SSN a été réutilisée pour la description des capteurs. Nous envisageons de plus de réutiliser l'ontologie modulaire du textile VetiVoc (Aimé et al., 2016), qui fournit un vocabulaire riche pour l'habillement.

Le scénario 8 (Restructuration des ressources ontologiques), a été appliqué pour l'organisation des modules. L'articulation entre ces modules dépend des scénarios d'exploitation et de fabrication d'un VI.

3 Exigences de modélisation et description du contexte

Nous considérons un VI comme « *un vêtement ordinaire, augmenté de composants électroniques capables de détecter des stimuli et d'adapter leur fonctionnement au contexte d'usage en changeant les traitements qu'ils embarquent* ». Ainsi, un VI remplit sa fonction principale en tant que vêtement et apporte une valeur ajoutée à l'utilisateur, comme par exemple

[2] https://www.w3.org/TR/ld-bp/
[3] http://www.neon-project.org/

capturer les postures et les gestes d'un utilisateur ou monitorer ses signes vitaux, etc. Ainsi, la fabrication et l'exploitation d'un VI sont dirigées par l'utilisation faite du vêtement et les besoins exprimés par les utilisateurs finaux (militaire, pompier, sportif, etc.). Pour cela, nous avons déterminé trois entités : « Vêtement », « Composant électronique » et « Domaine d'application », affectant la modélisation d'un VI (cf. Figure 1).

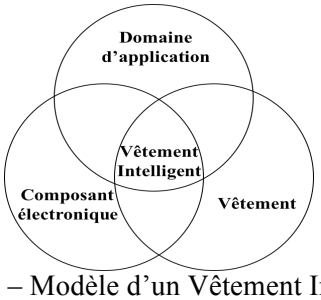


FIGURE 1 – Modèle d'un Vêtement Intelligent

« Vêtement », présente les caractéristiques des vêtements ordinaires (taille, genre, etc.) et définit une classification des vêtements (veste, cuissard, etc.).

« Composant électronique », modélise l'électronique embarquée dans le vêtement. Un composant électronique est lui-même constitué d'autres composants électroniques de niveaux de granularité différents (capteurs, microcontrôleur, etc.).

« Domaine d'application », définit le contexte d'utilisation pour lequel le vêtement sera fabriqué et exploité. Il peut être lui-même découpé selon la description du domaine.

Dans le cas du projet SmartSensing, les capteurs intégrés dans les vêtements sont des capteurs intelligents. Contrairement aux capteurs traditionnels, dits basiques, ils sont multifonctionnels (Akyildiz et al., 2012) et sont conçus pour être exploités dans plusieurs activités sportives. Un capteur intelligent intègre un ensemble de capteurs basiques et un microcontrôleur. Ce dernier peut exécuter différents algorithmes produisant des données calculées (indicateurs) à partir de mesures obtenues par les capteurs basiques. On peut adapter le capteur intelligent aux besoins des utilisateurs en changeant les algorithmes qui s'y exécutent. Ces besoins sont spécifiés selon le sport et ses pratiques ; chaque sport a un ensemble de pratiques et à chaque pratique est associé un ensemble d'indicateurs. Le tableau 1 montre un exemple pour le sport « Football ». D'un autre côté, nous associons, pour chaque sport, une panoplie vestimentaire (vêtements et optionnellement des accessoires).

TABLE 1 – *Exemple d'informations à retenir pour le sport « Football ».*

Pratique	Indicateurs	Panoplie
Compétition (exp. Match)	Température du corps, Niveau de stress, Géolocalisation	Maillot, cuissard, chaussettes
Musculation	Température du corps, Puissance	

4 Les modules de l'ontologie SMS

SMS comporte 6 modules relatifs aux différents regroupements de connaissances définis ci-dessus. Chaque module est associé à un domaine. Ces modules sont :

Module « 3SN » (*Semantic Smart Sensor Network*). Ce module décrit les différents capteurs intelligents déployés dans un vêtement de sport. Il est fondé sur une spécialisation de l'ontologie SSN. Il permet la découverte de capteurs et leur configuration pendant l'exploitation d'un VI. Il aide, aussi, pour la conception de capteurs intelligents lors de la fabrication d'un VI.

Module « Datasheet ». Ce module est une spécification des fiches techniques des capteurs basiques et offre une classification de ces derniers. Lors de la conception d'un nouveau capteur intelligent, ce module sert à sélectionner les meilleurs capteurs basiques.

Module « Measurement and Indicator ». Ce module décrit les algorithmes à coder et à déployer sur les capteurs intelligents en fonction du sport choisi. Il décrit principalement les

dépendances fonctionnelles et temporelles des données (dataflow) de ces algorithmes. Ces dépendances déterminent pour chaque indicateur les entrées nécessaires à son calcul.

Module « Sports & Activity ». Ce module définit les activités sportives. Il relie, pour chaque sport, l'ensemble de ses indicateurs et associe, à chaque type de sport, le vêtement adéquat.

Module « Clothing ». Ce module décrit les différentes panoplies de vêtements et d'accessoires de sport. Le lien de cette ontologie avec l'ontologie 3SN permet de spécifier les capteurs intelligents déployés et leur position.

Module « Sportsperson Profile ». Ce module concerne les caractéristiques des sportifs (taille, poids, âge, etc.) et permet d'associer à chaque sportif sa (ou ses) panoplie(s). De plus, il fournit la synthèse de ses données actimétriques pour chaque sport pratiqué.

5 Construction et structuration des modules pour la phase d'exploitation

L'exploitation d'un VI de sport, nécessite l'utilisation d'un ensemble de connaissances permettant de spécifier, selon la pratique du sport sélectionné, les indicateurs à calculer et ainsi les capteurs intelligents à configurer. La conceptualisation des modules mis en œuvre est fondée sur la spécialisation de l'ontologie Dolce UL qui se situe au niveau d'abstraction le plus élevé dans ces modules. Nous ne traitons, ici, que trois modules sur les six, sollicités dans la phase d'exploitation.

5.1 Modules spécifiques pour l'exploitation

«3SN». Pour représenter les capteurs intelligents en termes d'algorithmes, de microcontrôleur et de spécifications des capteurs basiques embarqués, nous réutilisons l'ontologie SSN, tout en respectant sa structure générique et son alignement avec DOLCE UL (cf. Figure 2 ; le préfixe «ssn» dénote les concepts de SSN et «sms3sn» ceux de 3SN). Le concept «sms3sn:SmartSensor» représente un capteur intelligent ; c'est une spécialisation de «ssn:SensingDevice». Il est composé de «sms3sn:BasicSensor» et de «sms3sn:Microcontrollor». Il possède des caractéristiques telles que sa dimension et son poids. Le microcontrôleur, représenté par le concept «sms3sn:Microcontrollor», exécute un ensemble d'algorithmes («sms3sn:Algorithm»), et dispose de capacités de traitement.

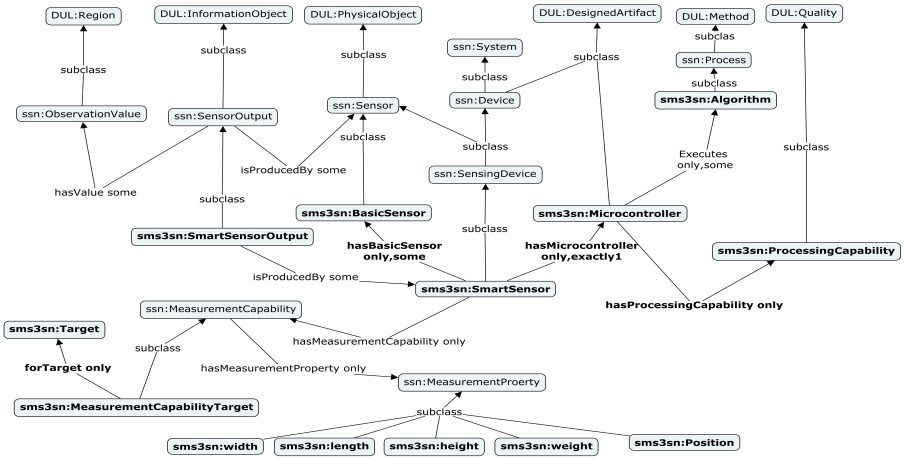


FIGURE 2 – Module 3SN

« Measurement and Indicator ». Ce module est représenté par la Figure 3. Les mesures et les indicateurs sont décrits comme des sous concepts de «smsmi:Data». Une classification des mesures et des indicateurs peut être créée comme une spécialisation de «smsmi:Measurement» et de «smsmi:Indicator». Pour créer le lien entre une instance d'une Data et une instance du

capteur qui la produit, une relation d'équivalence est établie entre «smsmi:Data» de ce module et « sms3sn:SmartSensorOutput » du module 3SN. Les dépendances fonctionnelles des données sont définies par les relations entre les inputs et les outputs des algorithmes.

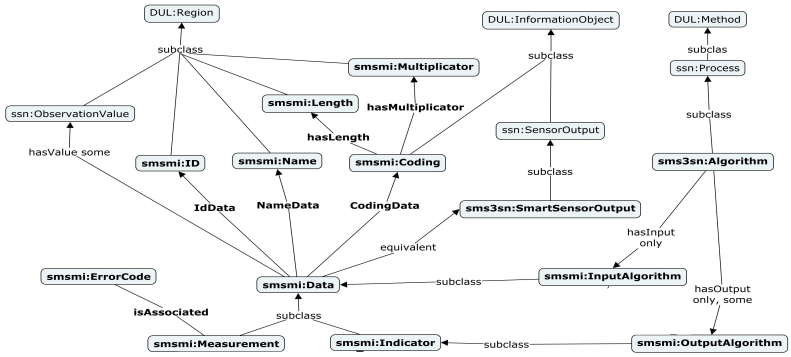


FIGURE 3 – Module « Measurement and Indicator »

« Sport and Activity ». Ce module (représenté dans la Figure 4) spécifie, de façon générique, les connaissances relatives aux activités sportives. Il permet de relier pour chaque sport l'ensemble de ses pratiques et de ses indicateurs. Il offre un modèle pour référencer les sports et décrit les contraintes qui leurs sont associées. Une classification des sports peut étendre ce module par l'ajout des différentes catégories de sports (Running, Tennis, Football, etc.). Nous ne présentons pas dans la Figure 4 ces catégories de sports. Ces dernières peuvent être ajoutées comme une spécialisation des concepts « smssa:CollectiveSport » ou « smssa:IndividualSport ». Elles peuvent être également ajoutées comme des instances d'un de ces concepts.

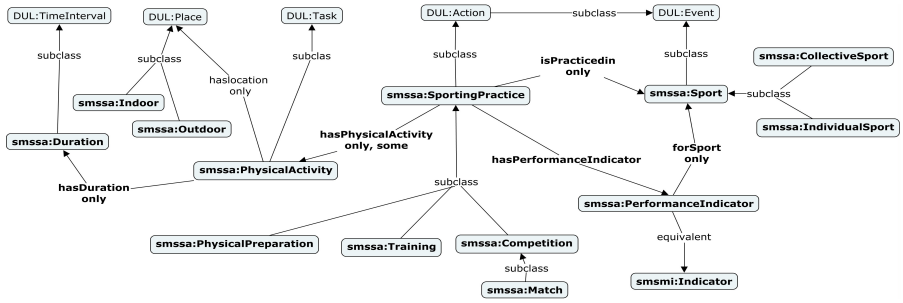


FIGURE 4 – Module « Sport and Activity »

« smssa:SportingPractice » modélise des instances de pratique. Il en existe plusieurs comme l'entraînement, la compétition, la récupération, etc. Dans la Figure 4, nous ne présentons que certaines pratiques. Pour les enrichir, il est possible de spécialiser le concept «sms:SportingPractice » . Une même pratique, par exemple l'entraînement, peut-être associée à plusieurs sports mais elle aura des indicateurs de performance spécifiques à chaque sport. Ainsi, les indicateurs de performance sont associés à une pratique dans le cadre d'un sport spécifique (via l'utilisation de l'axiome OWL « propertyChainAxiom » caractérisant la propriété « isPracticedin »). Ce module est associé au module « Measurement and Indicator » par le lien d'équivalence entre « smssa:PerformanceIndicator » et « smsmi:Indicator ».

5.2 Caractéristiques des modules

La Figure 5 montre la structure globale de SMS, créée à partir de liens, définis manuellement, entre les différents modules. Ces liens représentent les interconnexions permettant d'interroger et de raisonner sur l'ontologie SMS. Les modules publiés en ligne[4],

[4] http://3s-web.enstb.org/ontologies/SMS/

présentés dans la sous-section 5.1 sont une modélisation générique de chaque domaine en décrivant le fonctionnement et les contraintes associées. Leur extension, par spécialisation (dans le but de classification : d'indicateurs, de capteurs, de pratique ou de sport) peut être réalisée. Une synthèse métrique (nombre de concepts 'C' et de relations 'R') est donnée pour ces modules (triangle).

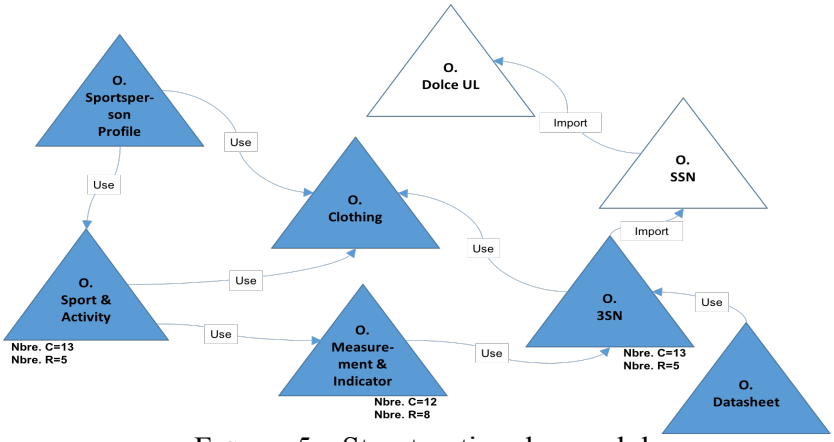


FIGURE 5 – Structuration des modules

6 Conclusion

Dans cet article, nous avons présenté l'ontologie modulaire SMS pour l'exploitation et la fabrication de vêtements intelligents. Ses modules décrivent les différents domaines d'un vêtement intelligent de sport. Ils sont fondés sur la réutilisation de ressources ontologiques de références. Basé sur un cadre méthodologique, décrit dans cet article, nous avons présenté les modules mis en œuvre à l'exploitation des vêtements intelligents et leur structuration en termes de relations sémantiques dans l'ontologie SMS. La validation de ces modules est en cours de réalisation par le développement d'un prototype du système visé.

Références

AIMÉ X., GEORGE S., HORNUNG J. (2016). VETIVOC : A modular ontology for the Fashion, Textile, and Clothing domain. In *Applied Oontology*, 11(1), pp.1–28.

AKYILDIZ, I. F., SU W., SANKARASUBRAMANIAM Y. & CAYIRCI, E. (2012). A Survey on Sensor Networks.

BIZER C., HEATH T & BERNERS-LEE T. (2009). Linked Data - The Story So Far. In *International Journal on Semantic Web and Information Systems,* 5(3), pp. 1-22.

BORST, W. N. (1997). Construction of engineering ontologies for knowledge sharing and reuse. These in University of Twente.

COMPTON, M., BARNAGHI, P., CORCHO, O., COX, S., HAUSWIRTH, M., HENSON, C., JANOWICZ, K., LEFORT, L., SHETH, A. & TAYLOR, K. (2012). The SSN Ontology of the W3C Semantic Sensor Network Incubator Group. In *Journal of Web Semantics*.

D'AQUIN M., SCHLICHT A., STUCKENSCHMIDT H. & SABOU M. (2007). Ontology modularization for knowledge selection: Experiments and evaluations. In Database and Expert Systems Applications, pp. 874-883.

RANTANEN J., IMPIO J., KARINSALO T., MALMIVAARA M., REHO A., TASANEN M. & VANHALA J. (2002). Smart Clothing Prototype for the Arctic Environment. Personal and Ubiquitous Computing. 6(1), pp. 3-16.

STECHER, R., NIEDEREE, C., NEJDL, W. & BOUQUET, P. (2008). Adaptive ontology reuse: finding and re-using sub-ontologies. In *International Journal of Web Information Systems*. 4(2), pp. 198-214.

SUAREZ-FIGUEROA, M. C., GOMEZ-PEREZ, A., & FERNANDEZ-LOPEZ, M. (2012). The NeOn Methodology for Ontology Engineering. In *Ontology Engineering in a Networked World*, pp. 9-34.

Un modèle pour la représentation des connaissances temporelles dans les documents historiques

Sahar Aljalbout, Gilles Falquet

Centre Universitaire d'informatique, Université de Genève
saharjalbout@gmail.com
Gilles.falquet@unige.ch

Résumé : Traiter et publier les données des sciences historiques dans le web sémantique constitue un défi intéressant où la représentation des aspects temporels joue un rôle clé. Nous proposons dans cet article un modèle de représentation des connaissances temporelles adapté au travail sur les documents historiques. Ce modèle est basé sur la notion de fluent que l'on représente dans des graphes RDF. Nous montrons comment ce modèle permet de représenter les connaissances nécessaires aux historiens et de raisonner sur celles-ci à l'aide des langages SWRL et SPARQL. Ce modèle est en cours d'utilisation dans un projet de numérisation, d'étude et de publication des manuscrits du linguiste Ferdinand de Saussure.

Mots-clés : web sémantique, représentation des connaissances temporelles, raisonnement temporel, documents historiques, ontologies historiques

1 Introduction

La représentation des connaissances temporelles avec les langages du web sémantique reste encore un défi, en particulier en terme de simplicité d'utilisation et de capacité de raisonnement. Pourtant, les sciences historiques et les données qu'elles produisent constituent un vaste domaine d'application dans lequel de nouvelles techniques de représentation, de raisonnement et de publication sur le web sémantique restent à créer.

Dans cet article nous présentons un modèle pour la représentation et le raisonnement temporel dans l'analyse des documents historiques. Ce modèle est actuellement appliqué dans le cadre d'un projet interdisciplinaire de publication savante des manuscrits de Ferdinand de Saussure. Saussure (1857-1913) est un linguiste suisse connu comme l'un des fondateurs de la linguistique moderne. De son vivant, Saussure n'a que très peu publié ses travaux. Il a par contre laissé de nombreux textes manuscrits (représentant plus de 30 000 pages) qui constituent une ressource d'une très grande richesse dans le domaine de la linguistique et de son histoire. Cependant leur étude est rendue complexe par le fait que nombre de ces textes ne sont pas datés et que la terminologie utilisée (et créée) par Saussure change considérablement au cours du temps. D'où l'intérêt de disposer d'un modèle de connaissances et de techniques d'inférence temporelles pour aider les experts saussuriens à indexer sémantiquement ces textes, à les dater ou à reconstituer leur séquence temporelle et en fin de compte à les comprendre.

En outre, la communauté des humanités numériques porte un grand intérêt à la publication des connaissances sur le web sémantique. C'est pourquoi nous avons décidé d'utiliser les technologies du web sémantique comme cadre d'implémentation du modèle.

2 Les besoins pour une représentation adéquate des connaissances historiques

Nous avons effectué une analyse détaillée des besoins des historiens de la linguistique concernant la dimension temporelle d'une base de connaissance. Le résultat de cette analyse a permis de classer les besoins de modélisation temporelle liée aux documents historiques selon quatre axes:

Représentation des entités contextuelles présentes dans les documents. La détermination des personnages, lieux, évènements auxquels les documents font référence permet d'exploiter le contexte historique pour mieux comprendre les documents.

Représentation des changements dans le contexte historique au cours du temps. Des propriétés telles que le lieu de résidence, la fonction ou les liens de collaboration d'une personne évoluent au cours du temps.

Représentation de la cause des changements (actions, évènements, ...)

Représentation des différentes terminologies utilisées dans les documents. La compréhension du contenu d'un document historique est toujours relative. Elle dépend de la terminologie employée par l'auteur et de la conscience qu'a le lecteur de l'usage de cette terminologie.

La prise en comte de ces besoins a guidé la construction et la validation de la partie structurelle de notre modèle.

3 Etat de l'art

La prise en compte du temps dans les bases de connaissance, et en particulier dans le Web sémantique a donné lieu à de nombreux travaux, dont nous ne mentionnerions que les plus pertinents pour notre étude.

3.1 Développements théoriques

Dans (Gutierrez et al., 2007) puis (Motik, 2012) les auteurs traitent du temps de validité en RDF et OWL. Les graphes temporels qu'ils proposent contiennent des étiquettes de validité temporelle associées à chaque triplet. Ils définissent une notion de conséquence logique entre ces graphes et des opérations d'interrogation

Les travaux sur la logique de description temporelle (Lutz et Wolter, 2008) introduisent quand à eux des opérateurs qui permettent de définir des concepts intrinsèquement temporels (p.ex. le concept d'*être mortel*) ..

Il existe également des approches tel *Event Calculus* (Mueller, 2008) qui ont pour but le raisonnement sur les actions et les changements qu'elles provoquent.

Un autre axe de recherche consiste à utiliser la notion de version d'ontologies (Klein et Fensel, 2001) pour représenter l'évolution de la connaissance (et de la terminologie).

3.2 Implémentation dans les langages existants

D'un point de vue plus pratique, certaines techniques ont été proposées pour utiliser les langages et systèmes existants pour la modélisation temporelle. Elles essaient en général de surmonter les limitations imposées par le seul usage des prédicats binaires en RDF et OWL. Parmi celles-ci on trouve :

- les *named graphs* (Tappolet et Bernstein, 2009) où chaque graphe contient les triplets qui sont vrais pour un intervalle de temps spécifié

- l'utilisation de patrons pour la représentation de relations n-aires en OWL et RDF(S) (Noy et Rector, 2006). La dimension temporelle peut alors être ajoutée à chaque relation binaire.
- les 4D-fluents (Welty et Fikes, 2006) où chaque concept temporel est représenté comme un objet à 4 "dimensions" avec comme quatrième dimension le temps.

3.3 Discussion

Les travaux présentés précédemment sont de natures différentes, riches et variés mais on constate que les travaux théoriques n'ont pas abouti à des langages et systèmes largement acceptés ou utilisés. Les solutions pragmatiques souffrent quand à elles de divers défauts allant de la prolifération d'objets dans les graphes à l'absence de mécanismes de raisonnement. Elles ne prennent pas non plus en considération tous les besoins de modélisation historique définis dans la section 2..

4 Un modèle temporel pour la représentation des connaissances historiques

Le modèle que nous avons défini pour représenter les connaissances liées aux manuscrits historiques comprend la représentation des manuscrits eux-mêmes (images, transcriptions, annotations, etc.), la représentation du contexte historique (personnes, lieux, évènements, terminologies employées, etc.) et les liens de référence entre manuscrits et entités du contexte.

4.1 Le modèle temporel et sa réalisation dans le Web sémantique

Pour représenter les changements dans le contexte historique et leurs causes nous proposons d'utiliser la notion de fluent (McCarthy et Hayes, 1969), (Mueller, 2008). Un fluent (propositionnel) est défini comme une fonction qui associe une valeur de vérité à un énoncé pour un instant donné.

Nous appellerons *assertion de fluent* un énoncé qui indique qu'un fluent est vrai pendant un intervalle de temps donné. Par exemple, l'énoncé « *Saussure a vécu en Allemagne entre 1876 et 1881* », indique que le fluent « *Saussure vit en Allemagne* » est vrai pendant toute la période 1876-1881. Par contre un énoncé tel que « *Saussure est né en 1857* », bien que temporel, ne concerne pas un fluent car la relation *est né en* entre Saussure et 1857 n'est pas susceptible de changer au cours du temps.

Le problème de la représentation des fluents dans le Web sémantique est lié aux langages de représentation tels que RDF qui ne supportent que les relations binaires. Même si les relations qui varient au cours du temps sont binaires, la représentation de l'intervalle temporel durant lequel la relation est vraie nécessite un troisième argument ou une réification. D'autre part, il est également nécessaire de pouvoir représenter la cause qui a initialement rendu vrai le fluent et celle qui l'a rendu faux à la fin de l'intervalle (quand elles sont connues).

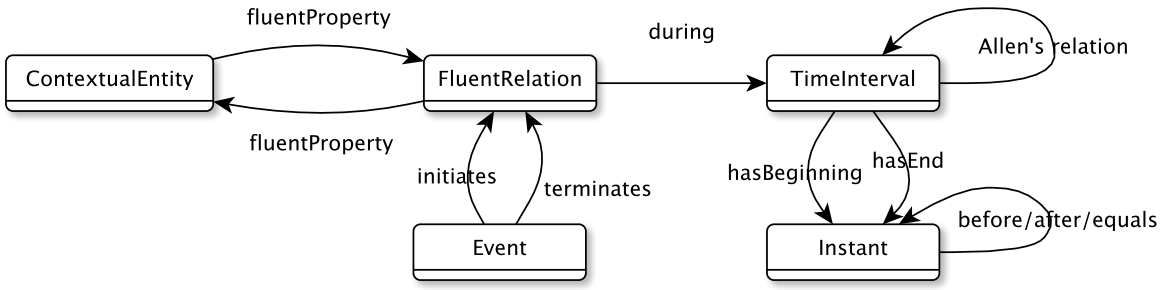


FIGURE 1 – Diagramme de classes RDFS de la représentation des fluents

Pour représenter une assertion de fluent nous utilisons un patron de relations *n-aires* (Noy et Rector, 2006) en introduisant un objet de type *FluentRelation* inspiré de (Preventis et al, 2012)), comme décrit sur la figure 1.On lui attribue le prédicat *during* et l'objet *TimeInterval* pour modéliser l'intervalle de temps durant lequel le fluent est valide. Un évènement peut initier ou terminer la période de validité (propriétés *initiates* et *terminates*). L'expression de l'intervalle de validité peut être quantitative, si l'on précise les instants de début et de fin, ou qualitative si l'on spécifie l'intervalle par ses relations de Allen (during, overlaps, meets, ...) avec d'autres intervalles.

La figure 2 montre la représentation de l'énoncé « Saussure a vécu en Allemagne entre 1876 et 1881 et en France entre 1881 et 1891 », qui contient deux assertions de fluents. Les entités contextuelles représentées dans ce schéma sont Saussure (une instance de la classe Personne) et Allemagne et France (deux instances de la classe Lieu). L'évènement qui a déclenché le premier fluent est le début des études universitaires de Saussure. Et l'évènement qui le pousse à changer son lieu de résidence est son enseignement et ses études à l'EPHE.

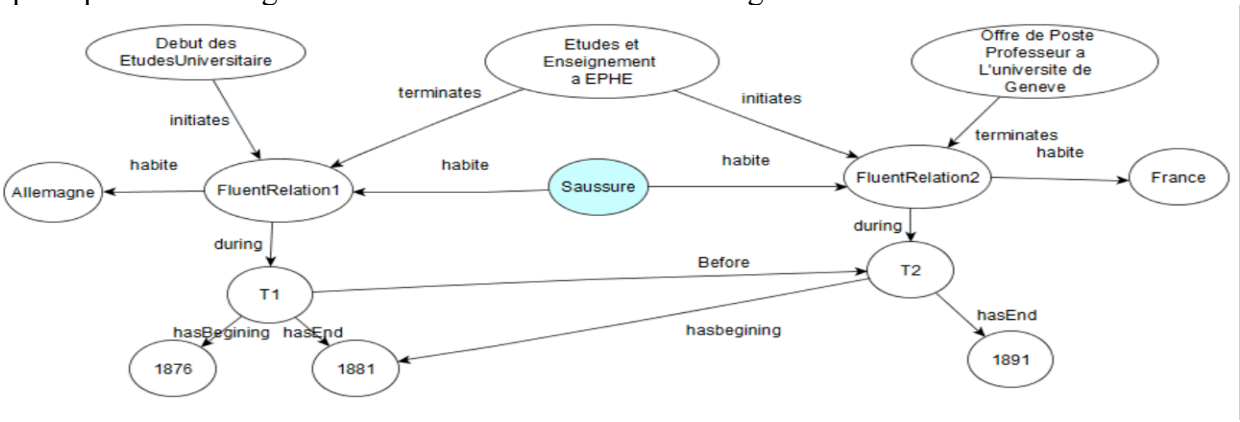


FIGURE 2 – Représentation RDF de deux assertions de fluents. Les types des objets sont omis

4.2 Modélisation de l'évolution terminologique

L'un des besoins importants des historiens (des sciences) concerne la représentation de la terminologique utilisée dans chaque document. En effet, dans le cas d'un auteur scientifique, les changements terminologiques sont fréquents car celui-ci travaille généralement sur des concepts encore instables ou crée lui-même de nouveaux concepts. C'est pourquoi, outre les entités contextuelles de type personne, évènement, lieu, relation scientifique, ... le modèle permet de représenter les terminologies utilisées par le ou les auteurs des documents. Chaque terminologie est formée d'entités de type concept associée à des termes et définitions.

La propriété *uses* permet de créer des assertions de fluents pour représenter le fait qu'un auteur utilise une terminologie particulière pendant une période (il n'est pas exclu qu'il en utilise plusieurs en parallèle !)

5 Raisonnement temporel sur le modèle

Dans cette section nous montrons comment le raisonnement temporel peut être réalisé pratiquement sur le modèle avec SWRL et SPARQL.

5.3 Inférences Temporelles avec SWRL et SPARQL

Il est possible d'utiliser directement des règles SWRL pour inférer des faits qui découlent des connaissances temporelles (fluents). On peut par exemple définir une règle temporelle

pour chaque propriété fonctionnelle: pour une propriété (fluente) fonctionnelle `prop` on n'a qu'une valeur pour un objet à un instant donné. Ce qui peut se traduire par la règle SWRL

```
prop(?x, ?f1, ?y1)[?i1], prop(?x, ?f2, ?y2)[?i2], overlaps(?i1, ?i2)
-> sameAs(?y1,?y2)
```

où `P(?X, ?F, ?Y)[?I]` est une abréviation de

```
FluentRelation(?F), P(?X,?F), P(?F,?Y), during(?F,?I)
```

Toutefois, les règles SWRL ne permettent pas la génération de nouveaux objets (noeuds RDF). Elle ne peuvent donc pas servir à créer de nouvelles assertions de fluents qui nécessitent la création d'instances des classes *FluentRelation* et *TimeInterval*.. Par contre, on peut utiliser des opérations SPARQL INSERT pour créer ces objets sous forme de nœuds blancs. Un exemple typique est la règle

Si un manuscrit M *écrit par* A *est une lettre à* B *et le temps d'écriture de* M *est* [t_1 .. t_2] *alors* A *connaît* B *pendant l'intervalle* [t_1 .. fin de la période considérée]

que l'on pourrait écrire en "SWRL étendu" sous la forme

```
Lettre(?l), auteur(?l, ?a), destinataire(?l, ?b), dateEcriture(?l, ?t1)
-> connait(?a, ?f, ?b)[?i], start(?i, ?t1), stop(?i, fin_periode)
```

et qui se traduit directement en SPARQL par

```
INSERT {?A :connait
         [a :FluentRelation ;
          :during [a :TimeInterval ; :start ?t1; :stop :fin_periode];
          :connait ?B]}
WHERE {?L a :Lettre. ?L :auteur ?A. ?L :destinataire ?B. ?L :
       dateEcriture ?t1}
```

On peut alors obtenir un système d'inférence complet en itérant l'exécution des règles SWRL et des insertions SPARQL de manière exhaustive.

Pour garantir que le processus se termine il faut cependant ajouter deux conditions

1. ne pas générer de nouveaux fluents superflus (qui ont les mêmes propriétés qu'un fluent déjà existant). Cette condition peut être incorporée directement dans les expressions d'insertion SPARQL sous forme d'un filtre.
2. les règles d'inférence (en SWRL étendu) ne doivent pas faire référence à des individus de *FluentRelation* ou de *TimeInterval* en position de sujet ou d'objet d'une assertion de fluent.

Ces conditions sont suffisantes car on ne considère qu'un ensemble fini d'instants (temps discret) et aucun autre type d'objet que des *FluentRelations* et *TimeInterval* n'est créé.

5.4 Construction de snapshots[1] pour le raisonnement synchronique

Afin de permettre le raisonnement synchronique nous avons introduit la notion de snapshot. Un *snapshot* sur un intervalle temporel [t_1, t_2] représente les faits qui restent vrais entre t_1 et t_2, c'est-à-dire tous les faits statiques et tous les fluents dont l'intervalle de validité contient [t_1, t_2]. Le snapshot est un graphe non temporel qu'on obtient en remplaçant les fluents valides sur cet intervalle par des triples non temporels, selon la règle

[1] Nous utiliserons le terme anglais plutôt que le terme *instantané* qui est d'usage peu courant.

```
prop(?x, ?f, ?y)[i], contains(i,[t1,t2])  -> prop(?x,?y)
```

qui s'implémente facilement par une opération SPARQL DELETE/INSERT.

Il devient alors possible d'interroger "synchroniquement" le snapshot avec des requêtes SPARQL ne faisant pas intervenir le temps ou d'appliquer des règles SWRL non temporelles.

Les faits inférés dans un snapshot sur $[t_1, t_2]$ peuvent ensuite être réinjectés dans la base de connaissance globale sous forme de fluents valides sur $[t_1, t_2]$. .

6 Conclusion

Les ressources du domaine historique constituent un objet d'étude intéressant pour les communautés du web sémantique. Nous avons proposé dans ce papier un modèle temporel historique et son implémentation avec les techniques du web sémantique. Ce modèle met l'accent sur la cause des changements au cours du temps et sur l'inférence de nouvelles connaissances temporelles. La partie structurelle du modèle a été testé sur de nombreux cas représentatifs fournis par les experts saussuriens. Il reste nécessaire de tester des techniques de raisonnement temporel, qui nécessiteront probablement des travaux d'optimisation. D'autre part nous devrons travailler avec les utilisateurs au recensement et à l'écriture des règles d'inférence temporelle les plus pertinentes dans le domaine étudié, ce qui permettra d'évaluer l'utilisabilité du modèle pour des humanistes.

Références

ALLEN J. F.(1983). "Maintaining Knowledge about Temporal Intervals." Commun. ACM, vol. 26, no. 11. p. 832–843.

BAADER F. & HOOROCKS I. & SATTLER U (2002). Description Logics for the Semantic Web. KI.16. p.57-59

GUTTIERREZ C. & HURTADO C. A. & VAISMAN A. A. (2007). Introducing Time into RDF. IEEE transactions on Knowledge and Data Engineering 19 (2). p. 207–218.

KLEIN M. & FENSEL D. (2001). Ontology versioning on the Semantic Web. Proceedings of SWWS'01, The first Semantic Web Working Symposium. p. 75-91.U SA, California, Stanford University.

LUTZ C. & WOLTER F. (2008). Temporal Description Logics.15th International Sympsium on Temporal Representation and Reasoning. p. 3-14. Canada, Universite du Quebec à Montreal.

MCCARTHY J. & HAYES P.J.(1969). Some Philosophical Problems from the Standpoint of Artificial Intelligence'. In D. Michie (ed), Machine Intelligence: 4. New York: Elsevier.

MOTIK B.(2012). Representing and querying validity time in RDF and OWL: A logic based approach. In Journal of Web Semantics: Science, Services and Agents on the World Wide Web. p. 3-21.

MUELLER E.(2008). Event Calculus. In Handbook of Knowledge representation. Foundations of Artificial Intelligence. p. 671-708.

NOY N. & RECTOR A. (2006). Defining N-ary Relations on the Semantic Web. [Online]. available:http://www.w3.org/TR/swbp-n-aryRelations/

PREVENTIS A. &MARKI P. & PETRAKIS G.M EURIPIDES, BATSAKIS S. (2014) Chronos : A tool for handling temporal ontologies in Protege. International Journal of Artificial Intelligence Tools.

TAPPOLET J. & BERNSTEIN A. (2009). Applied Temporal RDF: Efficient Temporal Querying of RDF Data with SPARQL. In: Proceedings of the European Semantic WebConference. p. 308-322

WELTY C. & FIKES R. (2006). A Reusable Ontology for Fluents in OWL. Formal Ontology in Information Systems, Proceedings of the Fourth International Conference. p. 226–236

Assister l'utilisateur à expliciter un modèle de trace avec l'analyse de concepts formels

Béatrice Fuchs

Univ Lyon, Université Lyon 3, CNRS , LIRIS, UMR5205, F-69372, France. beatrice.fuchs@liris.cnrs.fr

Résumé : Nous proposons d'analyser des jeux de données représentant des traces afin de découvrir le modèle sous-jacent et assister leur intégration dans un système à base de traces dédié au stockage et à la manipulation de traces numériques. Ce principe est mis en oeuvre dans CSV2KTBS, un prototype interactif qui s'appuie sur l'analyse de concepts formels pour générer un modèle de la trace et assister un utilisateur dans le processus d'intégration dans un système à base de traces. L'utilisateur est sollicité pour interpréter les concepts qui lui sont proposés à partir de l'analyse des traces, ainsi que pour valider d'autres résultats obtenus par l'analyse. Ces principes ont été appliqués aux traces du jeu sérieux Tamagocours pour l'apprentissage des règles juridiques de diffusion de ressources numériques ainsi qu'aux traces de ClassCraft, une application ludique pour l'enseignement secondaire.

Mots-clés : Traces, système à base de traces, extraction de connaissances, analyse de concepts formels, modélisation de traces

1 Introduction

Les traces témoignent d'une activité passée et sont précieuses pour étudier et comprendre des phénomènes complexes. Parfois elles constituent le seul matériau disponible sur lequel on peut s'appuyer pour inférer des connaissances et construire un modèle du phénomène étudié. Plus précisément les *traces numériques d'interaction* sont issues de l'observation de l'activité d'un utilisateur lorsqu'il réalise une tâche assistée par un système informatique. Actuellement de plus en plus d'outils sont disponibles pour la gestion de traces permettant leur production, leur collecte et leur exploitation en aval pour des usages variés (analyse, visualisation, raisonnement, découverte de connaissances). Une des premières étapes en vue de l'exploitation des traces est la collecte : les traces doivent d'abord être recueillies à partir d'une source, l'environnement informatique où l'utilisateur réalise une activité, ou un fichier. Puis elles doivent être traitées afin d'être exploitables à des fins d'analyse ou de raisonnement par exemple. De ce fait, il est préférable de les associer à un modèle de sorte que les différents éléments qui les composent soient compréhensibles et que les résultats des diverses analyses réalisées en aval de la collecte soient interprétables. Généralement, les traces sont fournies après le déroulement de l'activité et se présentent sous différentes formes (fichiers texte, csv, tsv, *etc.*). Dans la plupart des cas elles ne sont pas associées à modèle explicite. L'utilisateur qui prend en charge leur intégration dans un environnement d'exploitation des traces, même s'il les connaît bien, ne pourra s'affranchir d'une étape de modélisation qui peut être difficile selon la complexité des traces dont il dispose.

Nous nous intéressons dans cet article à l'assistance à l'intégration de trace dans une base de traces qui explicite de façon automatique un modèle à partir des données qui lui sont fournies, en interaction avec l'utilisateur. Nous proposons une méthode d'analyse interactive à l'aide de l'analyse de concepts formels afin d'expliciter le modèle sous-jacent d'une trace à partir d'un fichier de données brut de type csv. Nous proposons une mise en oeuvre avec les traces de Tamagocours, un jeu sérieux pour l'apprentissage de règles juridiques de diffusion de documents

numériques à des fins pédagogiques.

Dans le premier paragraphe nous présentons les principes d'un système à base de traces modélisées, la difficulté de la collecte, et le cadre du jeu Tamagocours sur lequel nous nous sommes appuyés. Nous détaillons le processus de génération du modèle de trace sur cet exemple. Nous discutons des limites de cette approche puis nous concluons avec quelques perspectives.

2 Traces d'interaction

Un Système à Base de Traces modélisées (SBT) est un système dédié à la modélisation et la manipulation générique de traces numériques (Champin *et al.*, 2013). Les traces sont collectées en enregistrant les actions de l'utilisateur sous la forme d'un ensemble d'actions datées et situées et appelées *éléments observés* ou *obsels* pour *Observed Elements*. Dans un SBT, les traces sont associées à un modèle constituant un premier niveau d'interprétation de la trace. Le modèle de trace décrit les différents *types d'obsel* associés à un ensemble d'*attributs* et de *relations* entre types d'obsel. Un système dédié à la gestion de traces permet de collecter des traces, les stocker et les manipuler à l'aide d'opérations génériques appelées transformations qui sont de différents types : filtrage d'obsels, fusion de traces, etc. Nous utilisons la plateforme kTBS [1] qui réifie cette notion de SBT. Dans kTBS, les traces et leurs modèles sont décrits à l'aide du formalisme RDF et de ce fait nous utilisons par la suite le format TURTLE pour décrire le modèle de trace généré.

2.1 Collecte de traces

Lors de l'intégration de traces dans un kTBS, la première étape consiste à décrire son modèle à partir des informations disponibles sur les traces dont on dispose. L'alimentation de la base de traces ne pose pas de problème lorsque tous les types d'obsel sont caractérisés par les mêmes attributs, mais ce n'est généralement pas le cas. Les types d'obsel sont souvent différenciés ce qui rend plus difficile le processus d'intégration, d'autant plus que l'on ne dispose très souvent que de peu d'informations sur la structure des différents types d'obsel composant la trace. La construction du modèle nécessite alors d'étudier la trace pour répertorier d'abord quels sont les actions représentées dans la trace, et pour chacune d'elles, inventorier les attributs qui les décrivent. Or, les formats de fichiers sont souvent variés et conçus ad-hoc pour les besoins spécifiques d'une application particulière. Dans (Bouvier *et al.*, 2014) par exemple, les traces sont issues d'un jeu en ligne et se présentent sous la forme d'un fichier csv dont les premières lignes sont consacrées à une description très sommaire des différents types d'obsel associés aux attributs les caractérisant. Une colonne est susceptible de représenter différents attributs selon le type d'obsel. Dans l'approche de (Besnaci *et al.*, 2015), l'intégration de multiples sources hétérogènes est réalisée par l'utilisateur qui décrit d'abord son modèle dans le kTBS et le met en correspondance avec les éléments des traces à importer. Dans ces approches, la construction du modèle reste manuelle à la charge de l'utilisateur. Or, dès que le nombre d'attributs et de types d'obsel augmente, la conception du modèle est souvent fastidieuse. De plus les traces peuvent contenir des anomalies qui rendent difficile la conception du modèle et sont susceptibles d'avoir un impact non seulement sur le modèle mais aussi sur les analyses ultérieures des traces. Pour assister l'utilisateur dans le processus d'intégration d'une trace dans un kTBS, nous

1. kernel for Trace Based System

proposons d'exploiter l'analyse de concepts formels (ACF). L'ACF a été largement utilisée dans de nombreux domaines, en ingénierie des ontologies ou en génie logiciel et couvre un large spectre d'applications (Poelmans *et al.*, 2013). Dans le paragraphe suivant nous décrivons le cas d'étude sur lequel nous nous appuyons pour présenter notre approche.

2.2 Tamagocours

Tamagocours (Sanchez *et al.*, 2015) est un jeu destiné à l'apprentissage des règles juridiques auxquelles est soumis l'usage de ressources numériques dans le contexte éducatif. Il a été conçu à l'intention d'étudiants destinés à la carrière d'enseignant dans le cadre de la mise en place du C2I2e [2]. Les utilisateurs sont répartis en équipes de 2 à 4 joueurs et doivent alimenter un «Tamago» en ressources pédagogiques numériques (images, sons, vidéos, livres, articles, publications conçues à des fins pédagogiques, partitions, etc.). Les ressources sont disposées sur une étagère où les utilisateurs peuvent consulter leurs caractéristiques (nature, date de parution, taille de l'extrait, droits d'auteur, etc.), les récupérer et les associer à un mode d'utilisation (présentation orale, projection en classe, sujet d'examen, mise en ligne sur l'intranet, etc.). Les caractéristiques des ressources conjointement à leur mode d'utilisation déterminent si l'utilisation des ressources est autorisée ou non. Les utilisateurs peuvent ensuite placer des ressources dans un réfrigérateur commun à l'équipe où elles restent consultables de la même manière que sur l'étagère, puis les donner au Tamago pour le «nourrir». Le Tamago est associé à un score qui évolue au fur et à mesure des réussites (ressource autorisée) ou échecs (utilisation d'une ressource hors du cadre légal) des actions des utilisateurs du groupe. Une succession d'échecs peut mener au dépérissement du Tamago. Les actions d'un utilisateur sont visibles par l'équipe, et le jeu comporte 5 niveaux de difficulté croissante. Pour mener à bien leur activité, les utilisateurs ont à leur disposition une aide qui comporte l'ensemble des règles juridiques applicables, ainsi qu'une messagerie instantanée afin de dialoguer entre eux et se concerter.

2.3 Les traces

Les traces de 244 utilisateurs ont été collectées dans un fichier au format csv qui représente au total 25 944 actions de jeu. Un extrait de ce fichier est montré dans la table 1. Le fichier, tel qu'il a été reçu initialement, comportait 13 types d'action différents décrits à l'aide de 24 attributs. Les types d'action possibles sont décrits dans la table 2. La collecte initiale comportait l'attribut `logType` représentant le type d'action utilisateur qui a été ensuite combiné à des valeurs d'autres d'attributs pour obtenir l'attribut `actionType` (combinaison de `feedTamago` avec `isWon` pour obtenir `feedTamagoGood` et `feedTamagoBad`. D'autres attributs résultent de combinaisons tels que `resourceTypeMoU` (combinaison de `resourceType` et `mode_of_use`) et `grpus_id` (combinaison de `group_id` et `user_id`).

Pour procéder à l'importation de ces traces dans un système à base de traces, il a d'abord fallu s'intéresser à la construction du modèle, c'est à dire définir l'ensemble des types d'obsel associés chacun à un ensemble d'attributs, et organiser les types d'obsel en hiérarchie d'héritage.

2. Certificat Informatique et Internet niveau 2 enseignant

id	date	actionType	group_id	user_id	grpus_id	Codage	message	help	res_id	item_id	resourceType	mode_of_use	resource_title	creationDate	rightsAgreements	res_size	item_size	reason	game_id	level_id	isWon
5	30/03/2015 11:05:45	fill Cupboard	2													/			2	1	1
7	30/03/2015 11:06:45	help Link	2	3	2_3			Accès aide en ligne								/			2	1	1
8	30/03/2015 11:06:49	tuto	2	3	2_3											/			2	1	1
9	30/03/2015 11:07:00	showItem CUPBOARD	2	3	2_3				119	552	journal		Le Nouvel	1964	Droits d'auteur CFC	/	10 articles		2	1	1
12	30/03/2015 11:07:42	showItem CUPBOARD	2	3	2_3				43	339	book		Le Grand M	1913	Domaine Public	/	4		2	1	1
16	30/03/2015 11:07:52	showItem CUPBOARD	2	3	2_3				113	529	journal		Le Figaro		Domaine public	/	intégrale		2	1	1
17	30/03/2015 11:07:56	showItem CUPBOARD	2	3	2_3				42	326	book		La littéra	2013	Droits d'auteur CFC	/419	1		2	1	1
18	30/03/2015 11:07:59	showItem CUPBOARD	2	3	2_3				43	335	book		Le Grand M	1913	Domaine Public	/	intégrale		2	1	1
19	30/03/2015 11:08:02	showItem CUPBOARD	2	4	2_4				108	508	journal		The Washin	1983	Droits d'auteur	/	5 articles		2	1	1
20	30/03/2015 11:08:19	addTo Fridge	2	3	2_3				43	335	book	printedCopies	Le Grand M	1913	Domaine Public	/	intégrale		2	1	1
21	30/03/2015 11:08:20	showItem CUPBOARD	2	4	2_4				117	541	journal		Science et	1913	Droits d'auteur	/	1 article		2	1	1
22	30/03/2015 11:08:38	chat	2	3	2_3	OJ	Quelqu'un sait ce qu'il faut faire??									/			2	1	1
23	30/03/2015 11:08:39	showItem FRIDGE	2	4	2_4				43	335	book	printedCopies	Le Grand M	1913	Domaine Public	/	intégrale		2	1	1
24	30/03/2015 11:08:45	chat	2	6	2_6	OJ	aucune idée!									/			2	1	1
25	30/03/2015 11:08:45	tuto	2	3	2_3											/			2	1	1
26	30/03/2015 11:09:04	feedTamago Good	2	3	2_3				43	335	book	printedCopies	Le Grand M	1913	Domaine Public	/	intégrale		2	1	1
27	30/03/2015 11:09:25	chat	2	3	2_3	OJ	Je mets des trucs dans le frigo et je lui file à bouffer									/			2	1	1
28	30/03/2015 11:09:30	showItem CUPBOARD	2	4	2_4				108	506	journal		The Washin	1983	Droits d'auteur	/	1 article		2	1	1
...																					
29175	09/04/2015 15:38:42	showItem CUPBOARD	89	177	89_177				407	63	image		Le baiser	1950	Droits d'auteur	/	72 dpi		731	5	1
29177	09/04/2015 15:39:14	addTo Fridge	89	177	89_177				223	686	video	assessments	Le bon la	1966	Droits d'auteur	161 m/	6 min		731	5	1
29178	09/04/2015 15:39:16	feedTamago Good	89	177	89_177				223	686	video	assessments	Le bon la	1966	Droits d'auteur	161 m/	6 min		731	5	1

TABLE 1 – Un extrait du fichier des traces du jeu Tamagocours. Les attributs sont représentés en colonnes, et la colonne `actionType` représente le type d'action réalisé par un utilisateur. Chaque ligne est une action repérée par un identifiant (`id`) et par une estampille (`date`).

3 Génération interactive du modèle de trace

On peut remarquer que la table 1 comporte des intersections vides : pour certains types d'action, il n'y a aucune valeur pour certains attributs. Cela signifie que ces attributs ne s'appliquent pas à ces types d'action. Ces «valeurs manquantes» permettent d'identifier à quels types d'action les attributs correspondants doivent être rattachés. Il est ensuite possible d'en déduire quels types d'action partagent quels attributs. C'est précisément l'objet de l'analyse de concepts formels.

3.1 L'analyse de concepts formels

L'analyse de concepts formels (ACF, (Ganter & Wille, 1999)) vise à représenter sous forme de treillis une relation binaire entre un ensemble d'objets et un ensemble de propriétés, décrite dans un tableau à deux dimensions appelé contexte formel. Elle repose sur une théorie mathématiques (Barbut & Monjardet, 1970) issue de la théorie des treillis. Typiquement, un contexte formel décrit la relation binaire «a pour propriété» entre un ensemble d'objets et un ensemble de propriétés. Un concept formel est défini par une paire (E, I) où E est un ensemble d'objets et I un ensemble d'attributs où :

addToFridge	ajouter une ressource dans le frigo
chat	envoyer un message
feedTamagoBad feedTamagoGood	Nourrir le Tamago avec une bonne ou une mauvaise ressource
fillCupboard	Initialisation de l'étagère par le logiciel avec un ensemble de ressources
helpLink	affichage de l'aide
removeFromFridge	supprimer une ressource du frigo
showItemCUPBOARD showItemFRIDGE, showItemLEVEL showItemTAMAGO showItemSTOMACH	examiner une ressource placée sur l'étagère examiner une ressource dans le frigo, examiner une ressource dans le tableau de fin de niveau, examiner une ressource dans le Tamago. examiner les ressources dans l'estomac du Tamago.
tuto	Consultation du tutoriel

TABLE 2 – Les différents types d'action utilisateur dans le jeu Tamagocours.

— E est appelé l'extension et contient l'ensemble des objets partageant les attributs de I,
— I est appelé l'intension et contient l'ensemble des attributs partagés par les objets de E.

Les concepts formels peuvent être ordonnés en hiérarchie sous la forme d'un treillis de concepts appelé encore treillis de Galois, où les relations d'ordre partiel entre les ensembles d'objets d'une part et les ensembles d'attributs d'autre part forment une correspondance de Galois. L'ACF extrait les concepts à partir d'un contexte formel qui représente une relation binaire entre les individus et les propriétés, présenté dans la section suivante.

3.2 Contexte formel

Il a donc fallu dans un premier temps répertorier la liste des types d'action de Tamagocours et déterminer pour chacune d'elles la liste des attributs pour lesquels ils sont définis. Le tableau 3 représente cette relation binaire «a pour attribut», entre les types d'action et les attributs : c'est un contexte formel.

Définition 1 (Contexte formel)
Soient deux ensembles finis O et A, et une relation $R \subseteq O \times A$.
Deux fonctions sont associées à la relation R :

— *La fonction f permet de connaître les attributs partagés par un ensemble d'objets :*
$f : \mathcal{P}(O) \to \mathcal{P}(A), X \mapsto f(X) = \{y \in A | \forall x \in X, (x, y) \in R\}$
— *La fonction g permet de connaître les objets partageant un ensemble d'attributs*
$g : \mathcal{P}(A) \to \mathcal{P}(O), Y \mapsto f(Y) = \{x \in O | \forall y \in Y, (x, y) \in R\}$

Plus précisément, O est l'ensemble des types d'action Tamagocours, et A l'ensemble des attributs utilisés pour les décrire. Le contexte formel est défini de la façon suivante : pour chaque type d'action $o_i \in O$, pour chaque attribut $a_j \in A$, soit $V_{i,j}$ l'ensemble des valeurs prises par les objets de type o_i pour l'attribut a_j, $(o_i, a_j) \in R \iff V_{i,j} \neq \emptyset$. Il s'agit donc de répertorier les couples $(a, o) \in A \times O$ pour lesquels il existe au moins une ligne possédant une valeur.

	showItem STOMACH	addTo Fridge	help Link	tuto	fill Cupboard	showItem CUPBOARD	showItem FRIDGE	showItem LEVEL	showItem TAMAGO	remove FromFridge	feedTamago Good	chat	feedTamago Bad
actionType	X	X	X	X	X	X	X	X	X	X	X	X	X
codage						X						X	X
creationDate	X	X				X	X	X	X	X	X		X
date	X	X	X	X	X	X	X	X	X	X	X	X	X
game_id	X	X	X	X	X	X	X	X	X	X	X	X	X
group_id	X	X	X	X	X	X	X	X	X	X	X	X	X
grpus_id	X	X	X	X	X	X	X	X	X	X	X	X	X
help			X										
id	X	X	X	X	X	X	X	X	X	X	X	X	X
isWon	X	X	X	X	X	X	X	X	X	X	X	X	X
item_id	X	X				X	X	X	X	X	X		X
item_size	X	X				X	X	X	X	X	X		X
level_id	X	X	X	X	X	X	X	X	X	X	X	X	X
logType	X	X	X	X	X	X	X	X	X	X	X	X	X
message												X	
mode_of_use	X	X					X	X	X	X	X		X
reason	X	X					X	X	X	X			X
resource_id	X	X				X	X	X	X	X	X		X
resource_size	X	X	X	X	X	X	X	X	X	X	X	X	X
resource_title	X	X				X	X	X	X	X	X		X
resourceType	X	X				X	X	X	X	X	X		X
resourceTypeMoU	X	X				X	X	X	X	X	X		X
rightsAgreements	X	X				X	X	X	X	X	X		X
user_id	X	X	X	X		X	X	X	X	X	X	X	X

TABLE 3 – La relation entre les types d'action et les attributs. Les colonnes représentent les types d'action et les lignes les attributs utilisés pour les décrire.

3.3 Concepts formels

A partir du contexte formel, il est possible de construire un treillis de concepts. Il s'agit de déterminer les sous-ensembles maximaux d'objets (l'extension, ici les types d'obsel) qui partagent des sous-ensembles maximaux d'attributs (l'intension).

Définition 2 (Concept formel)
Un concept formel C *est un couple* (E, I) *tel que* $f(E) = I$ *ou de façon équivalente,* $E = g(I)$.
$E = \{o \in O | \forall a \in A, (o, a) \in R\}$ *est appelé l'extension du concept,*
$I = \{a \in A | \forall o \in o, (o, a) \in R\}$ *est appelé l'intension du concept.*

La figure 1 montre le treillis obtenu sur les traces à partir du contexte formel de la table 3. Les nœuds sont étiquetés par trois types d'informations : un numéro de nœud, l'ensemble des types d'action, c'est-à-dire l'extension, puis l'ensemble des attributs partagés c'est-à-dire l'intension. Le nœud le plus haut étiqueté c0 souvent appelé «Top» correspond à l'ensemble des attributs partagés par tous les types d'action. Ces attributs sont «hérités» par tous les nœuds placés «en dessous» dans la hiérarchie (ils ne sont pas répétés dans la figure pour une meilleure lisibilité). Le nœud le plus bas étiqueté c10 souvent appelé «Bottom» correspond à l'ensemble des types d'action partageant tous les attributs. Dans la figure 1, il n'y en a aucun. De la même façon que pour les attributs, les types d'action sont «hérités» mais cette fois de façon ascendante dans les nœuds situés «au dessus» dans la hiérarchie. Sur le treillis de la figure 1 des éléments sont mis en évidence car il montrent des anomalies au niveau des concepts c0-c1, c5-c7 et c2-c4, c2-c6-c9 : ces concepts ne peuvent être interprétés et semblent ne pas avoir de sens.

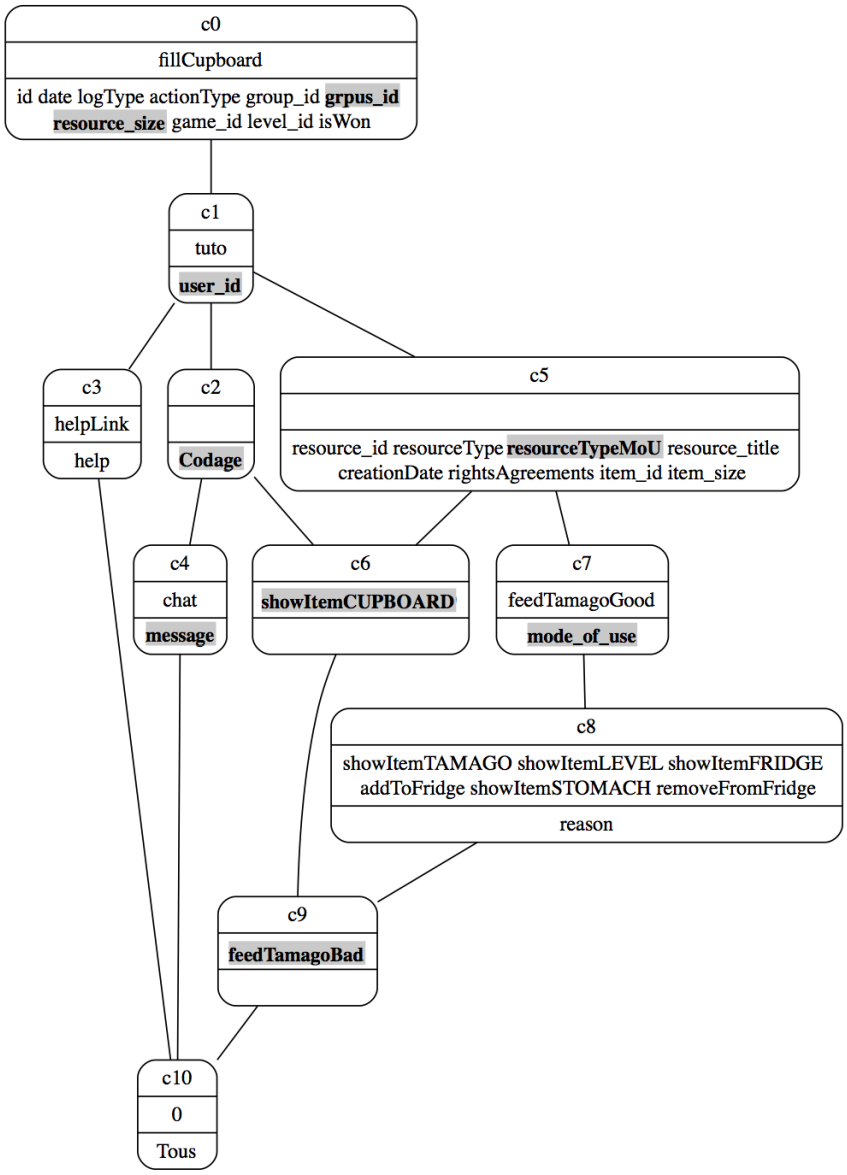


FIGURE 1 – Le treillis de concepts obtenu à partir du contexte formel de la table 3.

3.4 Rectification des données

Les modifications des données réalisées après la collecte ont généré des erreurs. C'est le cas en particulier pour les attributs `grpus_id`, `resourceTypeMoU` et `resource_size` :

— `grpus_id` est une concaténation de `group_id` et de `user_id`. L'action `fillCupboard` est générée lors du remplissage de l'étagère avec des ressources au début d'un jeu. Elle n'est pas associée à un utilisateur mais à un groupe et ne comporte donc pas de valeur pour `user_id`. La concaténation des deux attributs n'a pas pris la précaution de vérifier l'existence d'une valeur pour les deux attributs. Après vérification, 663 obsels possèdent une valeur pour `grpus_id` mais pas pour `user_id` :

ce sont toutes les actions de type `fillCupboard`. On retrouve la même anomalie avec l'attribut `resourceTypeMoU` qui résulte de la concaténation de `resourceType` et `mode_of_use`. Toutes les actions non associées à l'attribut `mode_of_use` ont une valeur pour `resourceTypeMoU`, sans `mode_of_use`. Après vérification, on constate qu'il s'agit de 9522 actions de type `showItemCUPBOARD`. Enfin de la même façon, `resource_size` ne se retrouve pas associé aux attributs liés à une ressource dans le concept pour la même raison.

— Les chats sont délicats à traiter car ils sont peu contextualisés : on ne peut pas connaître la ressource sur laquelle ils portent, ou leur objet. Pour cette raison un codage manuel a été réalisé afin de préciser leur objet. Compte tenu du nombre important d'actions dans la trace, des erreurs ont pu être introduites. D'après la figure 1, et après vérification, il s'avère que deux actions `feedTamagoGood` et `showItemCUPBOARD` ont été enrichies par erreur d'un champ codage qui a été introduit sur une mauvaise ligne, d'où la présence des concepts c6 et c7 qui n'ont pas lieu d'exister.

A l'issue de la correction des anomalies dans les données, on obtient le treillis de la figure 2.

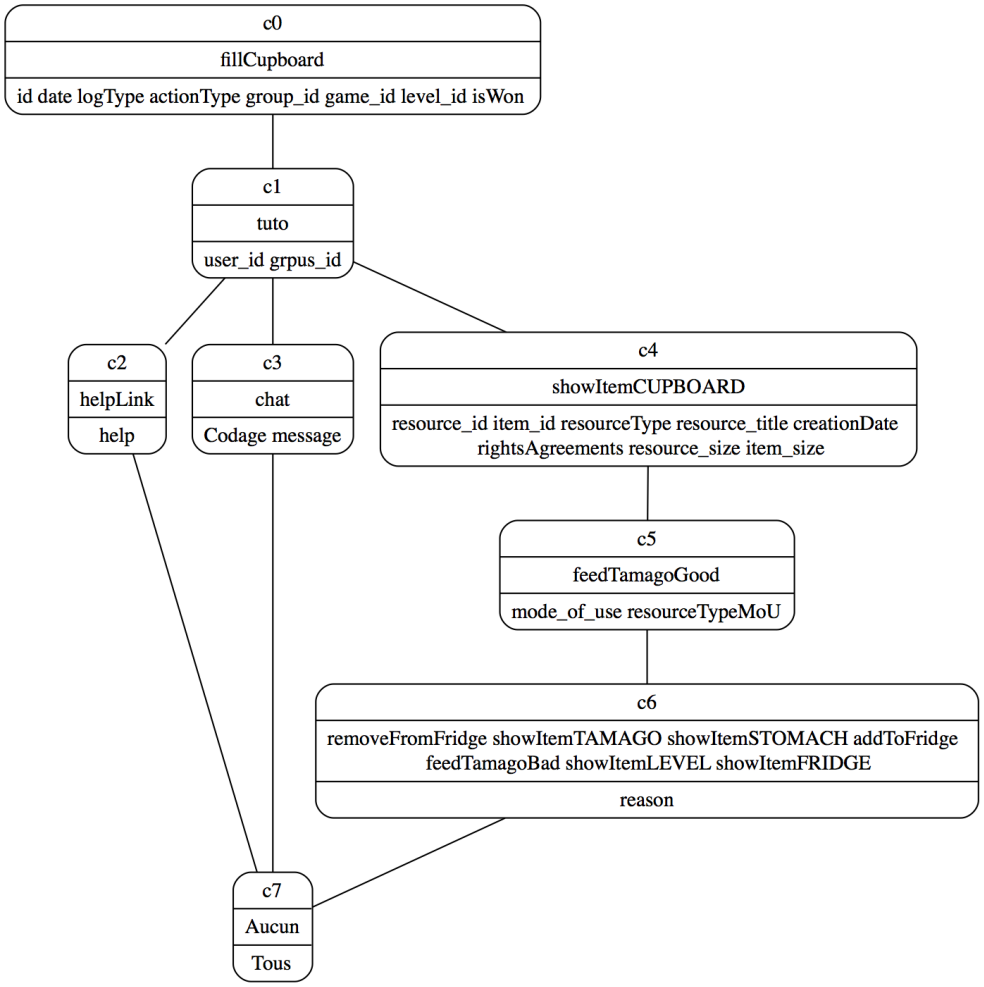


FIGURE 2 – Le treillis obtenu à l'issue de la correction des erreurs.

3.5 Interprétation du treillis

Après correction des erreurs il devient possible d'interpréter les différents concepts du treillis de la figure 2. On remarque qu'il est possible de supprimer le concept «Bottom» qui comporte tous les attributs mais aucun type d'action. On obtient la hiérarchie du modèle de trace de la figure 3.

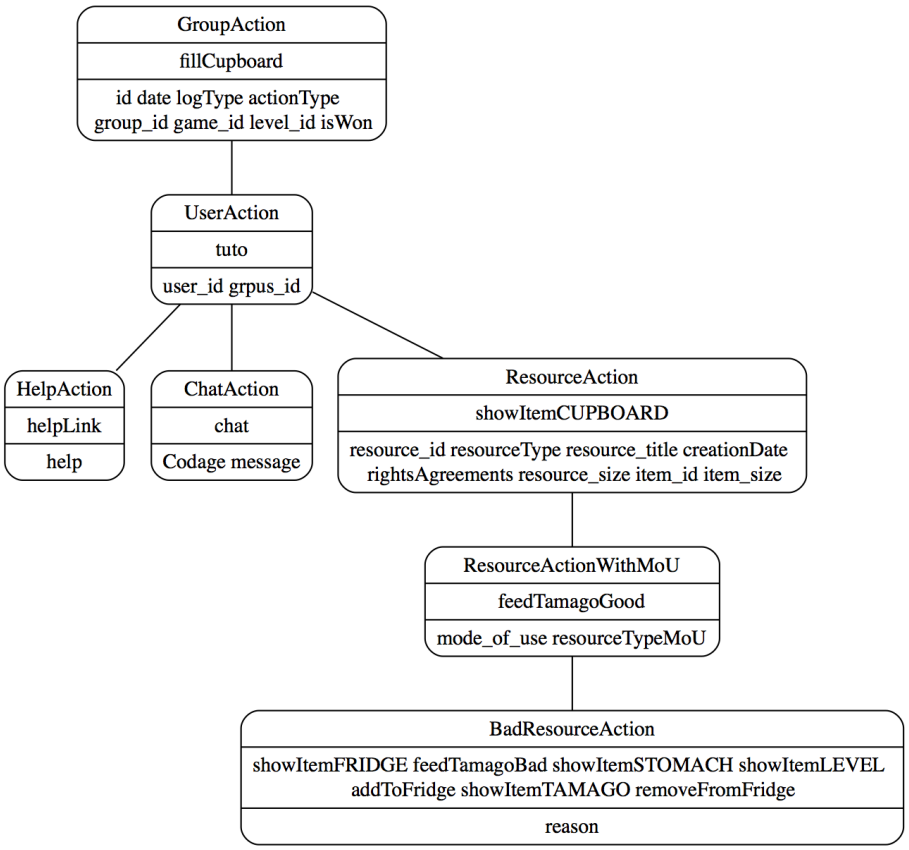


FIGURE 3 – Le modèle de trace obtenu à partir du treillis de concepts auquel l'élément «Bottom» a été supprimé et les concepts étiquetés.

Les concepts du treillis obtenu ont été interprétés de la façon suivante : Le type d'obsel `GroupAction` comporte les attributs qui sont partagés par toutes les actions, ce qui correspond au concept de groupe d'utilisateurs sur lequel le jeu est fondé. Le type d'obsel `UserAction` comporte les attributs correspondant à des actions d'utilisateurs, notamment l'identifiant de l'utilisateur. Le type d'obsel `ResourceAction` décrit les actions de manipulation des ressources sans mode d'utilisation. Le type d'obsel `ResourceActionWithMoU` décrit les actions de manipulation des ressources associées à un mode d'utilisation. Enfin le type d'obsel `BadResourceAction` représente les actions d'alimentation du Tamago avec des ressources et mode d'utilisation non autorisés et précise la raison de l'échec (`reason`).

Pour générer le modèle de trace, il suffit ensuite d'exploiter les informations présentes dans chaque concept du treillis de la figure 3 : les intensions et les extensions. Tout d'abord, les concepts formels sont utilisés pour générer des types d'obsel qui sont associés chacun à un

ensemble d'attributs (l'intension) partagé par les types d'action (l'extension). Les types d'action du jeu sont représentés par les types d'obsel qui héritent des définitions des types d'obsel issus des concepts formels ci-dessus. Par exemple le concept `UserAction` est représenté par un type d'obsel caractérisé par les attributs `user_id` et `grpus_id`, il hérite du type d'obsel `GroupAction` tous les attributs `id`, `date`, etc. Le type d'action `tuto` hérite du type d'obsel `UserAction`. Cette première étape de construction du modèle de trace permet d'obtenir un regroupement des attributs en types d'obsel et la liste des types d'obsel correspondant aux types d'action du jeu qui en héritent dans le modèle. Une deuxième étape précise le type et les caractéristiques des attributs et génère le modèle au format TURTLE pour le kTBS.

3.6 Génération du modèle de trace

Les données sont analysées de façon à déterminer les caractéristiques des valeurs prises par chaque attribut (liste des valeurs, intervalles de valeurs, etc.). En particulier pour chaque attribut, un type de données est détecté (nombre entier, nombre fractionnaire, booléen, chaîne de caractères, date et heure), et les attributs jouant le rôle d'identifiant sont détectés. Tous ces résultats sont proposés à l'utilisateur qui peut les valider ou les modifier. Cette analyse des types n'est pas détaillée davantage ici car elle sort du cadre de ce travail. Elle vise simplement à faciliter le travail de l'utilisateur. Ce processus est de plus perfectible, seuls les quelques types énumérés précédemment sont pris en compte. Ensuite, le modèle de trace est généré à l'aide de la syntaxe TURTLE. Un extrait du modèle de trace généré à partir du treillis de la figure 3 est présenté ci-dessous :

```
@prefix : <http://liris.cnrs.fr/silex/2009/ktbs#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
<http://localhost:8001/TamagocoursBase/> :contains <> .
<> a :TraceModel ;
    :hasUnit :second .
<#GroupAction> a :ObselType .
    <#id> a :AttributeType ;
        :hasAttributeDomain <#GroupAction> ;
        :hasAttributeRange xsd:integer.
    <#date> a :AttributeType ;
        :hasAttributeDomain <#GroupAction> ;
        :hasAttributeRange xsd:dateTime.
    <#actionType> a :AttributeType ;
        :hasAttributeDomain <#GroupAction> ;
        :hasAttributeRange xsd:string.
    <#group_id> a :AttributeType ;
        :hasAttributeDomain <#GroupAction> ;
        :hasAttributeRange xsd:integer.
    [...]
<#fillCupboard> a :ObselType ;
    :hasSuperObselType <#GroupAction> .
<#UserAction> a :ObselType ;
    :hasSuperObselType <#GroupAction> .
    <#user_id> a :AttributeType ;
        :hasAttributeDomain <#UserAction> ;
        :hasAttributeRange xsd:integer.
    <#grpus_id> a :AttributeType ;
        :hasAttributeDomain <#UserAction> ;
        :hasAttributeRange xsd:string.
<#tuto> a :ObselType ;
    :hasSuperObselType <#UserAction> .
[...]
```

Une fois le modèle de trace généré, l'importation de la trace peut être réalisée. Cette étape est opérationnelle et brièvement décrite ci-après. Elle repose sur l'analyse des attributs, en particulier les attributs de type date. Il s'agit de déterminer l'unité de temps à utiliser pour calculer les estampilles associées à chaque obsel, conformément aux unités de temps utilisées par le kTBS : seconde, milli-seconde, ou numéro séquentiel. La plus petite date est utilisée comme référence et les estampilles de chaque obsel sont calculées comme un nombre d'unités de temps depuis depuis la date de référence.

4 Discussion

Actuellement, la génération du modèle implémentée dans CSV2KTBS, ainsi que celle de la trace est entièrement opérationnelle, mais plusieurs points peuvent être améliorés. Il reste à développer une interface graphique pour améliorer l'interactivité, à connecter CSV2KTBS à un kTBS pour procéder à l'importation en «temps réel» du modèle et de la trace, et à prendre en compte les relations entre les obsels. CSV2KTBS est capable de traiter très rapidement une trace de presque 25 944 obsels. La construction du modèle ne souffre pas de la taille importante de la trace puisque c'est le nombre d'attributs et de types d'action qui sont prises en compte pour la construction du treillis de concepts, et ce nombre est relativement limité. Antérieurement, le travail de modélisation avait été réalisé «à la main» en élaborant manuellement un contexte formel à partir des données. Le modèle obtenu était très similaire mais comportait de nombreuses imperfections liées aux erreurs dans les données. L'intérêt de l'ACF s'est imposé de façon assez évidente et son utilisation a montré que l'ACF peut constituer un outil puissant d'assistance à l'explicitation d'une sémantique «cachée» dans des données, difficile et fastidieuse à réaliser manuellement dès lors que le nombre de types d'action et d'attributs dépasse une ou deux dizaines. Par ailleurs, l'explicitation d'une hiérarchie d'héritage des attributs présente un intérêt important pour les utilisations en aval de la trace à des fins d'analyse. Certaines méthodes de fouille de données peuvent tirer parti d'une telle hiérarchie, par exemple (Marinica *et al.*, 2008) ou (Brisson & Collard, 2008). La principale limitation de l'approche présentée dans cet article est qu'elle ne peut s'appliquer que si chaque colonne ne représente qu'un seul attribut. Ainsi, lorsqu'un type d'attribut n'est pas défini pour un type d'action, alors les intersections correspondantes dans le tableau de données sont vides, et c'est cette caractéristique de la trace qui permet de construire le contexte formel de la relation binaire. La façon dont les données tabulaires sont organisées dans les fichiers de types `csv` est très variable et cette approche n'est pas toujours applicable. Cette limitation a été constatée dans les traces du jeu sérieux ClassCraft [3], une application ludique destinée à favoriser l'apprentissage dans l'enseignement secondaire. La trace comportait 13 488 lignes correspondant aux actions réalisées par les utilisateurs avec 11 types d'action et 22 attributs. Le treillis a été généré correctement, mais le travail d'interprétation a été rendu difficile par le manque de documentation suffisante sur la signification des attributs et des types d'action. De plus, certaines colonnes représentent plusieurs attributs différents selon le type d'action. Afin de mieux capturer la signification de chaque attribut «caché», une phase supplémentaire en interaction avec l'utilisateur est nécessaire afin de différencier et expliciter les attributs selon le type d'action.

3. `https://www.classcraft.com`

5 Conclusion

Dans cet article nous présentons une approche pour la génération automatique et interactive d'un modèle de trace à partir de données au format csv à l'aide de l'analyse de concepts formels, mise en œuvre dans le prototype CSV2KTBS. L'approche a été appliquée sur les traces du jeu Tamagocours, ce qui a permis de mettre en lumière les anomalies contenues dans la trace et de contribuer à améliorer la qualité des données. Le modèle est généré d'abord sous forme graphique pour être plus facilement compréhensible par l'utilisateur. Après l'interprétation des concepts, le modèle de trace est généré au format TURTLE, puis la trace elle-même. L'approche est opérationnelle, générique et s'est avérée applicable sur d'autres traces. Il reste à développer une interface graphique pour améliorer l'interactivité du prototype, et à prendre en compte la différenciation des attributs lorsqu'ils prennent des significations différentes en fonction des types d'action. Une autre perspective est la conception d'un processus interactif qui permette la détection et la correction d'erreurs dans la source de données analysée, et la génération de requêtes en conséquence afin d'aider l'utilisateur à repérer ces erreurs et les corriger.

Références

BARBUT M. & MONJARDET B. (1970). *Ordre et classification : Algèbre et combinatoire.* Hachette, Paris.

BESNACI M., GUIN N. & CHAMPIN P.-A. (2015). Acquisition de connaissances pour importer des traces existantes dans un système de gestion de bases de traces. In *IC2015*, Rennes, France : AFIA.

BOUVIER P., SEHABA K. & LAVOUÉ E. (2014). A trace-based approach to identifying users' engagement and qualifying their engaged-behaviours in interactive systems : application to a social game. *User Modeling and User-Adapted Interaction*, **24**(5), 413–451.

BRISSON L. & COLLARD M. (2008). How to semantically enhance a data mining process ?. In *ICEIS*, volume 19, p. 103–116 : Springer.

CHAMPIN P.-A., MILLE A. & PRIÉ Y. (2013). Vers des traces numériques comme objets informatiques de premier niveau : une approche par les traces modélisées. *Intellectica*, (59), 171–204.

GANTER B. & WILLE R. (1999). *Formal Concept Analysis.* Springer, mathematical foundations edition.

MARINICA C., GUILLET F. & BRIAND H. (2008). Post-processing of discovered association rules using ontologies. In *Workshops Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*, p. 126–133.

POELMANS J., IGNATOV D. I., KUZNETSOV S. O. & DEDENE G. (2013). Formal concept analysis in knowledge processing : A survey on applications. *Expert systems with applications*, **40**(16), 6538–6560.

SANCHEZ E., EMIN-MARTINEZ V. & MANDRAN N. (2015). Jeu-game, jeu-play, vers une modélisation du jeu. Une étude empirique à partir des traces numériques d'interaction du jeu Tamagocours. *STICEF*, **22**.

Concepts de plus proches voisins dans des graphes de connaissances

Sébastien Ferré

IRISA/Université de Rennes 1
Campus de Beaulieu, 35042 Rennes cedex
ferre@irisa.fr

Résumé : Nous introduisons la notion de *concept de voisins* comme alternative à la notion de distance numérique dans le but d'identifier les objets les plus similaires à un objet requête, comme dans la méthode des plus proches voisins. Chaque concept de voisins est composé d'une intension qui décrit symboliquement ce que deux objets ont en commun et d'une extension qui englobe les objets qui se trouvent entre les deux. Nous définissons ces concepts de voisins pour des données complexes, les graphes de connaissances, où les nœuds jouent le rôle d'objets. Nous décrivons un algorithme *anytime* permettant d'affronter la complexité élevée de la tâche et nous présentons de premières expérimentations sur un graphe RDF de plus de 120.000 triplets.

Mots-clés : Graphes de connaissances, Web sémantique, Plus proches voisins, Analyse de concepts formels, Graph-FCA, Concepts de voisins, Hypergraphes, Homomorphismes.

1 Introduction

La méthode des k plus proches voisins (k-NN) (Mitchell, 1997) est à la fois simple et polyvalente. Elle permet la classification supervisée et le raisonnement à partir de cas (RàPC) (De Mantaras *et al.*, 2005). Son principe est, partant d'une instance, de trouver dans un jeu de données les instances les plus similaires, les *plus proches voisins*, et de s'appuyer sur ces voisins pour prédire la classe (classification) ou adapter une solution (RàPC). C'est une forme d'apprentissage paresseux (*lazy learning*) puisque les instances sont mémorisées tel quel, sans généralisation explicite.

Notre objectif est de pouvoir appliquer la méthode des k-NN aux graphes de connaissances, où chaque nœud est une instance et où le graphe entier participe à la description de chaque instance. Par exemple, dans les données de MONDIAL (May, 1999), les pays, continents, rivières, etc. sont mutuellement décrits les uns par les autres. Bisson (2000) distingue deux types de similarités. Les similarités *numériques* ont l'avantage d'être souples et économiques mais elles sont peu explicables et peuvent être trompeuses en cachant derrière une même valeur des similarités très différentes. De plus, il existe peu de telles mesures sur des données relationnelles (ex., RIBL (Horváth *et al.*, 2001)). Les similarités *symboliques* évitent ces inconvénients en produisant des représentations symboliques généralisant plusieurs instances. Néanmoins, à notre connaissances, ces similarités ont seulement été utilisées pour la formation de concepts et règles par généralisation (ex., PLI (Muggleton, 1995)), pas pour la recherche de plus proches voisins. De plus, les travaux existants considèrent généralement comme instances des petits graphes (ex. molécules (Kuznetsov, 2013)) plutôt que les nœuds d'un grand graphe. Un inconvénient des similarités symbolique est leur coût de calcul.

La contribution de ce papier est de proposer une approche à la fois symbolique, générale et efficace des plus proches voisins pour des graphes de connaissances (ex., graphes RDF, graphes

conceptuels (Chein & Mugnier, 2008)). Cette approche, CNN (*Concepts of Nearest Neighbours*), définit la similarité symbolique entre deux instances comme un *concept de voisins* dont l'intension représente ce que les deux instances ont en commun, et dont l'extension englobe les autres instances qui se trouvent "entre" les deux instances. La taille de l'extension peut tenir lieu de distance et la taille de l'intension de similarité numérique. Des clusters de plus proches voisins peuvent ainsi être identifiés et exploités comme dans l'approche classique. L'approche est purement *symbolique* parce que le graphe est exploité tel quel, sans recodage ou extraction de features, et parce que chaque plus proche voisin est accompagné d'une représentation intelligible de sa similarité (intension). L'approche est *générale* car elle couvre une large classe de graphes, les multi-hypergraphes, et ne requiert aucun paramètre (ex., pas de limite dans la profondeur d'exploration du graphe). Enfin, l'approche est *efficace* en focalisant les comparaisons sur l'instance requête et grâce à des techniques originales : *énumération de concepts par partitionnement* et *forme paresseuse de jointure* pour le calcul des extensions de concepts.

L'article est organisé comme suit. La section 2 situe notre approche par rapport aux approches existantes. La section 3 rappelle les définitions de Graph-FCA, le cadre dans lequel nos travaux se placent. La section 4 définit la notion de "concept de voisin" et la section 5 détaille un algorithme efficace pour les calculer. Enfin, la section 6 décrit nos premières expérimentations sur des données réelles, avant de conclure dans la section 7.

2 Travaux existants

La principale approche k-NN pour des données relationnelles est RIBL (*Relational Instance-Based Learning*) (Horváth *et al.*, 2001). Elle définit une distance numérique entre objets sur la base de l'exploration arborescente de l'hypergraphe à partir de cet objet jusqu'à une profondeur limite, sans prendre en compte les éventuels cycles. Nous avons appliqué une approche similaire pour guider des utilisateurs dans la production de descriptions RDF à partir d'exemples (Hermann *et al.*, 2012). À notre connaissance, aucune approche k-NN ne considère autre chose que des valeurs numériques comme représentation des distances entre objets.

Dans le domaine de l'Analyse de concepts formels (FCA) (Ganter & Wille, 1999), des notions proches de nos "concepts de voisins" ont été proposées mais elles s'appliquent à des données non-relationnelles. Kuznetsov (2013) utilise de tel concepts pour faire de la classification. Pour chaque exemple à classer, cela nécessite de calculer un concept pour chaque exemple déjà classé. Cette approche a été appliquée à des graphes mais où les objets sont de petits graphes (ex., molécules) et non pas les nœuds d'un grand graphe de connaissances. Nous avions proposé (Ferré & Ridoux, 2002) une approche similaire à celle de Kuznetsov mais avec un calcul des concepts de voisins dirigé par la description de l'objet à classer plutôt que par l'énumération des objets déjà classés.

Il existe deux extensions de la FCA applicables à des graphes de connaissances : RCA (*Relational Concept Analysis*) (Rouane-Hacene *et al.*, 2013) et Graph-FCA (Ferré, 2015). Nous avons choisi de nous appuyer formellement sur Graph-FCA car elle permet de définir chaque concept de voisins indépendamment des autres concepts. De plus, Graph-FCA autorise les hypergraphes (relations n-aires) et prend en compte les cycles, contrairement à RCA.

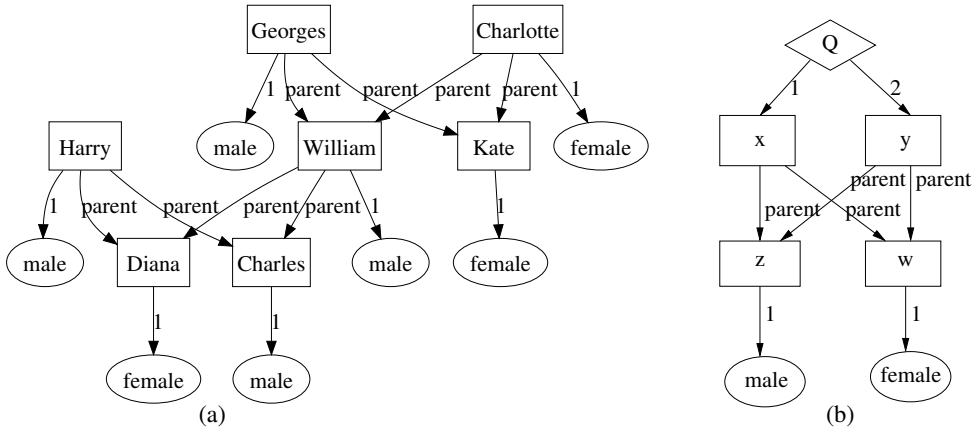


FIGURE 1 – (a) Contexte graphe sur la famille royale. Les rectangles sont les objets, les étiquettes d'arcs et d'ovales sont les attributs. (b) Un PGP définissant la relation binaire "frère-ou-sœur". Les rectangles sont des variables et le losange est le tuple de projection.

3 Graph-FCA : concepts formels d'un graphe de connaissances

Dans cette section, nous rappelons les principales définitions de Graph-FCA, une extension de l'analyse de concepts formels (FCA) (Ganter & Wille, 1999) aux graphes de connaissances, que nous avons introduite récemment (Ferré, 2015; Ferré & Cellier, 2016). Ces définitions sont illustrées par un petit exemple sur la famille royale anglaise.

Définition 1 (contexte graphe)
Un contexte graphe *est un triplet $K = (O, A, I)$, où O est l'ensemble des* objets, *A est l'ensemble des* attributs, *et $I \subseteq O^* \times A$ est une* relation d'incidence *entre des tuples d'objets (O^*) et des attributs.*

Un *contexte graphe* modélise un multi-hypergraphe orienté où les nœuds sont les *objets*, les étiquettes d'arcs sont les *attributs*, et les arcs sont des *incidences* $(\overline{o}, a)$ (aussi notée $a(\overline{o})$) entre tuples d'objets et attributs. L'aspect "orienté" vient de l'utilisation de tuples pour les arcs et l'aspect "hyper" vient de leur longueur quelconque. L'aspect "multi-" vient du fait qu'un même tuple peut être en incidence avec plusieurs attributs. La figure 1.(a) montre une représentation graphique d'un petit exemple. Des exemples d'arcs sont $male(William)$ et $parent(William, Diana)$. Différentes sortes de graphes de connaissances peuvent être traduit en contexte graphe : ex., graphes RDF, graphes conceptuels (Chein & Mugnier, 2008), contextes RCA (Rouane-Hacene *et al.*, 2013). Pour RDF, on a les traductions suivantes où o_x est l'objet représentant le nœud RDF x : triplets $\langle x \text{ a } c\rangle \mapsto a(o_x)$ (les classes deviennent des attributs unaires), autres triplets $\langle x\ p\ y\rangle \mapsto p(o_x, o_y)$ (les propriétés deviennent des attributs binaires), URI $u \mapsto u(o_u)$ et litéral $s \mapsto s(o_s)$ (les identités/valeurs deviennent des attributs unaires).

Définition 2 (pattern de graphe (projeté))
Un pattern de graphe *$P \subseteq \mathcal{V}^* \times A$ représente une généralisation d'un contexte graphe en abstrayant les nœuds-objets par des nœuds-variables pris dans un ensemble infini $\mathcal{V}$. Un* pattern de graphe projeté *(PGP) est un couple $Q = (\overline{x}, P)$ où P est un pattern de graphe et $\overline{x}$ est un tuple de projection sur certains nœuds-variables du pattern.*

Un PGP est analogue à une requête SPARQL de type SELECT-WHERE. La figure 1.(b) montre un PGP définissant la relation binaire "frère-ou-sœur" comme deux personnes partageant un père et une mère. Pour tout tuple d'objets $\overline{o} \in O$ d'un contexte $K = (O, A, I)$, l'expression $Q(\overline{o}) = (\overline{o}, I)$ dénote un PGP où les objets doivent être pris comme des variables et qui représente la description du tuple d'objets $\overline{o}$ par le graphe de connaissances K. Ainsi, dans l'exemple de la famille royale, chaque individu est décrit par le graphe entier, mais chacun d'un point de vue différent. Par exemple, William est un homme qui a un garçon et une fille qui ont tous deux une même mère, et qui a un père et une mère qui ont un fils en commun.

Les PGPs sont partiellement ordonnées par une relation d'inclusion : $Q_1 \subseteq_q Q_2$ ssi il existe un homomorphisme de graphe ϕ (Hahn & Tardif, 1997) du pattern de Q_1 vers le pattern de Q_2 qui préserve le tuple de projection. On note $dom(\phi)$ le domaine de définition d'un homomorphisme ϕ. On appelle *homset* Φ un ensemble d'homomorphismes ayant un même domaine. On note $homs_K(P)$ l'ensemble des homomorphismes d'un pattern P dans un contexte graphe K. La jointure de deux homsets est défini par $\Phi_1 \bowtie \Phi_2 = \{\phi_1 \cup \phi_2 \mid \phi_1 \in \Phi_1, \phi_2 \in \Phi_2, \phi_1 \sim \phi_2\}$, où $\phi_1 \sim \phi_2$ ssi pour tout $x \in dom(\phi_1) \cap dom(\phi_2)$, $\phi_1(x) = \phi_2(x)$. L'intersection de deux PGPs Q_1, Q_2 est défini par l'opérateur $\cap_q$, basé sur le produit de graphe dit catégorique (Hahn & Tardif, 1997), et donne le plus grand PGP qui soit inclu au sens de $\subseteq_q$ dans Q_1 et Q_2.

Définition 3 (concept graphe)
Soit $K = (O, A, I)$ un contexte graphe. Un concept graphe *de K est une paire (R, Q), constituée d'une relation (*extension*) et d'un PGP (*intension*), tel que $R = \{\overline{o} \mid Q \subseteq_q Q(\overline{o})\}$ et $Q \equiv_q \cap_q \{Q(\overline{o})\}_{\overline{o} \in R}$. L'*arité *d'un concept est la longueur du tuple de projection de son intension, et donc aussi la longueur des tuples de son extension.*

Les concepts sont définis de façon similaire à la FCA classique mais avec des PGP à la place des ensembles d'attributs et avec des *relations*, c'est-à-dire des ensembles de tuples d'objets, à la place d'ensembles d'objets. Un résultat important est que l'ensemble des concepts de même arité k forme un treillis $(\mathcal{C}_k, \leq, \wedge, \vee)$ où $C_1 \leq C_2$ signifie que C_1 est plus spécifique que C_2, c'est-à-dire a une plus petite extension et une plus grande intension. Dans ce papier, on se limite aux concepts unaires ($k = 1$) dont les extensions sont des ensembles d'objets, même si les résultats présentés s'appliquent également aux concept n-aires. Un exemple de concept est le couple $C_{ex} = (\{Charlotte, George, Harry, William\}, (x, \{parent(x, f), male(f), parent(x, m), female(m), parent(y, f), parent(y, m), male(y)\}))$. Il représente le concept des "enfants", lesquels, dans ce contexte, ont tous un père et une mère connus, lesquels ont toujours un fils.

4 Concepts de (plus proches) voisins

L'idée de départ est de remplacer la notion de distance numérique par une notion de distance conceptuelle.

Définition 4 (distance conceptuelle)
Soit $K = (O, A, I)$ un contexte graphe. La distance conceptuelle *entre deux objets $u, v \in O$ est le plus petit concept graphe unaire qui contient ces deux objets, c'est-à-dire le concept $\delta(u, v) = (R, Q)$ où le PGP $Q = Q(u) \cap_q Q(v)$ représente tout ce que u et v ont en commun et où la relation R englobe tous les objets qui partagent cette description commune.*

La distance conceptuelle vérifie les propriétés d'une distance si on prend l'ordre partiel $\leq$ sur les concepts et le supremum $\vee$ de concepts comme addition :

1. (positivité) $\delta(u,u) \leq \delta(u,v)$; ($\delta(u,u)$ joue le rôle de distance zéro)
2. (symétrie) $\delta(u,v) = \delta(v,u)$;
3. (inégalité triangulaire) $\delta(u,w) \leq \delta(u,v) \vee \delta(v,w)$.

On notera qu'une inégalité entre concepts $C_1 \leq C_2$ implique une inclusion entre leurs extensions $C_1.ext \subseteq C_2.ext$ et donc une inégalité entre le cardinal de ces extensions : $|C_1.ext| \leq |C_2.ext|$. La réciproque n'est pas vraie car $\leq$ est un ordre partiel sur les concepts. On peut donc utiliser les cardinaux d'extensions comme forme numérique – et dégradée – des distances conceptuelles : $d(u,v) := |\delta(u,v).ext|$. Cette forme numérique mesure en quelque sorte le nombre d'objets qui se trouvent entre u, v. On peut faire une observation similaire avec les intensions des concepts : $C_1 \leq C_2$ implique $|C_1.int| \geq |C_2.int|$. On peut donc utiliser la taille des intensions comme mesure de similarité : $sim(u,v) := |\delta(u,v).int|$.

Définition 5 (concepts de voisins)
Soit $K = (O, A, I)$ un contexte graphe et $u \in O$ un objet. Les concepts de voisins *de l'objet requête u (CN pour "concepts of neighbours") sont l'ensemble des distances conceptuelles partant de u, c'est-à-dire*

$$CN(u) := \{\delta(u,v) \mid v \in O\}.$$

L'extension propre d'un concept de voisins $\delta \in CN(u)$ est l'ensemble des objets qui se trouvent exactement à la distance δ de u : $\delta.proper := \{v \in O \mid \delta(u,v) = \delta\}$.

Le concept de voisins à distance zéro $\delta(u,u)$ englobe, en plus de u, les objets w dont la description contient celle de u ($Q(u) \subseteq_q Q(w)$). Comme il est inférieur à tous les autres concepts de voisins (positivité), on peut lui attribuer le rang 0, puis définir le rang de tout autre concept de voisins $\delta \in CN(u)$ par $rank(\delta) := 1 + max\{rank(\delta') \mid \delta' \in CN(u), \delta' < \delta\}$. On peut ensuite définir les concepts de plus proches voisins comme les concepts de rang 1, c'est-à-dire ceux qui sont immédiatement supérieurs au concept zéro. Un avantage de notre approche est de pouvoir définir l'ensemble des k plus proches voisins comme les instances de ces concepts de rang 1, sans avoir à fixer de valeur pour k. Un autre avantage est que ces plus proches voisins sont partitionnés en un ou plusieurs concepts, chacun apportant une justification via son intension.

Par exemple, le concept C_{ex} de la section précédente est la distance conceptuelle entre $Charlotte$ et $William$ (ils ont en commun d'avoir un père et une mère partageant un fils) et fait donc partie des concepts de voisins de chacun des deux individus. Pour $u = William$, l'extension propre est réduite à $\{Charlotte\}$ car $George$ et $Harry$ ont aussi en commun avec $William$ d'être des garçons. Pour $u = Charlotte$, l'extension propre est $\{Harry, William\}$ car $George$ a aussi en commun avec $Charlotte$ d'avoir des grands-parents dans le contexte graphe. Les k plus proches voisins de $William$ sont : (1) $Kate$ (parent d'un garçon et d'une fille), (2) $Charles$ (père), et (3) $Harry$ et $George$ (fils). Dans ce cas $k = 4$.

5 Algorithme

D'après les définitions de la section 4, l'algorithme naïf pour énumérer l'ensemble des concepts de voisins $CN(u)$ est le suivant.

Algorithme 1 (énumération naïve)
Require: *un graphe contexte* $K = (O, A, I)$*, un objet* $u \in O$
Ensure: *l'ensemble de concepts de voisins* CN

```
1: CN ← ∅
2: for all v ∈ O do
3:     Q ← Q(u) ∩_q Q(v)                          // calcul intension
4:     R ← {o ∈ O | Q ⊆_q Q(o)}                    // calcul extension
5:     CN ← CN ∪ {δ} where δ = (R, Q)
6:     δ.proper ← δ.proper ∪ {v}                  // v est dans l'extension propre de δ
7: end for
```

Dans le cas des graphes de connaissances de type Web sémantique, on doit supposer que le graphe a une seule composante connexe et donc que les descriptions des objets $Q(u), Q(v)$ sont de la taille du graphe. Le calcul de Q par l'intersection $\cap_q$ est donc quadratique dans le nombre d'arcs du contexte graphe, sans compter l'extraction du sous-pattern équivalent minimal (noyau) du pattern produit, ce qui est NP-complet. Le calcul de l'extension R implique un test d'inclusion $\subseteq_q$ NP-complet [1] pour chaque objet et donc pour chaque nœud du graphe. Le tout doit être calculé pour chaque objet.

Tout cela rend l'utilisation de cet algorithme naïf impratiquable et la complexité en NP-complet des opérations sur les graphes semble laisser peu d'espoir. Il est tentant de simplifier le problème en restreignant la description des objets à une certaine profondeur et en se ramenant à des arbres (pas de cycles), voire à des ensembles de chaines (pas de branchements). Cependant, cela signifirait que l'on ne travaille plus vraiment sur des graphes, mais sur des strucures plus pauvres (ex., arbres, chaines). Notre objectif est de calculer les concepts de voisins directement sur le graphe et sans restriction de profondeur ou autre. La nature des graphes de connaissances et des distances conceptuelles rend la tâche moins complexe qu'elle ne l'est dans le cas général :

1. les graphes de connaissances sont des (hyper)graphes *orientés* et *étiquetés*, ce qui pose des contraintes fortes sur les homomorphismes et réduit donc leur combinatoire ;
2. les descriptions d'objets et les intensions de concepts sont des PGPs, i.e. des graphes *centrés* sur un nœud qui sert d'ancrage et contraint encore davantage les homomorphismes.

Pour rendre pratiquable le calcul des concepts de voisins, nous combinons deux techniques originales. La première technique (section 2) consiste à énumérer ces concepts en partitionnant de façon de plus en plus fine l'ensemble des objets, plutôt que de traiter les objets un à un. Cela permet de grandement factoriser le calcul des intensions et extensions de concepts. La deuxième technique (section 3) consiste en une forme "paresseuse" de jointure entre homsets pour le calcul des extensions. Elle permet de représenter de façon compacte le homset d'un pattern dont la taille peut rapidement exploser.

5.1 Énumération par partitionnement

Le parti pris est de discriminer de plus en plus finement l'ensemble des objets par partitionnements successifs, en formant des sous-PGP de plus en plus grands de $Q(u)$. Un *pré-concept* matérialise une étape de ce processus avec une intension susceptible d'être agrandie et avec une extension propre susceptible d'être partitionnée davantage.

1. Il s'agit du problème de l'existence d'un homomorphisme entre deux graphes (Hahn & Tardif, 1997).

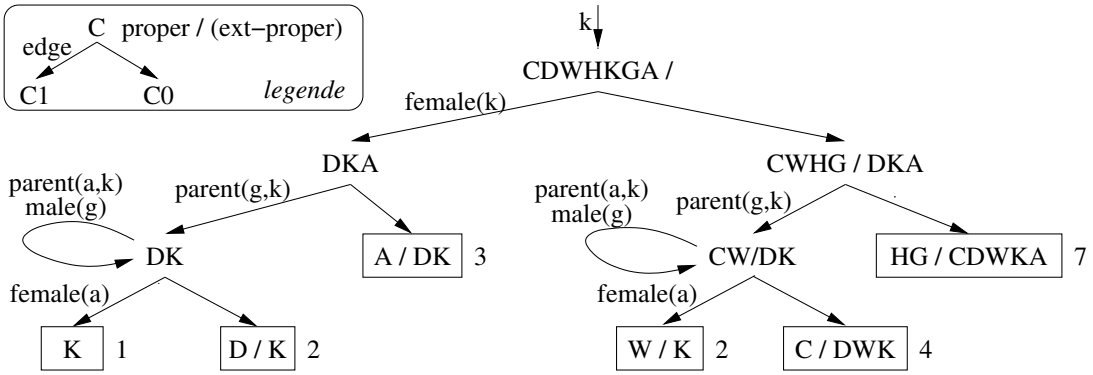


FIGURE 2 – Énumération par partition pour le contexte de la famille royale et avec $u = Kate$.

Définition 6 (pré-concept)
Un pré-concept *est une structure* $C = (pattern, homs, proper, edges)$*, où :*
- $pattern \subseteq I$ *est un sous-graphe connexe du contexte graphe contenant le nœud* u *;*
- $homs = homs_K(pattern)$ *représente l'ensemble des occurences du pattern dans* K *;*
- $proper \subseteq O$ *est l'extension propre du pré-concept ;*
- $edges \subseteq I$ *est l'ensemble des arcs restant disponibles pour partitionner le pré-concept.*

On définit l'extension et l'intension du pré-concept à partir du pattern et du homset en les projetant sur u *:* $int := (u, pattern)$ *;* $ext := \{\phi(u) \mid \phi \in homs\}$.

Le processus de partitionnement commence avec le pattern $\{\top(u)\}$, où $\top(u)$ est l'arc neutre qui ne pose aucune contrainte sur u, et l'ensemble O comme extension propre. Ensuite, chaque pré-concept peut être partitionné en deux pré-concepts par tout arc disponible et connexe au pattern. Le processus de partitionnement s'arrête pour un pré-concept quand son extension propre est vide ou bien quand il n'y a plus d'arcs disponibles.

Algorithme 2 (énumération par partitionnement)
Require: *un graphe contexte* $K = (O, A, I)$*, un objet requête* $u \in O$
Ensure: *un ensemble de pré-concepts de voisins* CN
1: $CN \leftarrow \{C_{init} := (\emptyset, homs_K(\{\top(u)\}), O, I)\}$
2: ***while*** *il existe* $C \in CN$ *et un arc* $e = (\overline{x}, a) \in C.edges$ *connexe avec* $C.pattern$ ***do***
3: $\quad C_1 = (C.pattern \cup \{e\},\ C.homs \bowtie homs_K(\{e\}),\ C.proper \cap C_1.ext,\ C.edges \setminus \{e\})$
4: $\quad C_0 = (C.pattern,\ C.homs,\ C.proper \setminus C_1.proper,\ C.edges \setminus \{e\})$
5: $\quad CN \leftarrow CN \setminus \{C\}$
6: $\quad$ **if** $C_1.proper \neq \emptyset$ **then** $CN \leftarrow CN \cup \{C_1\}$
7: $\quad$ **if** $C_0.proper \neq \emptyset$ **then** $CN \leftarrow CN \cup \{C_0\}$
8: ***end while***

La figure 2 illustre cet algorithme avec le calcul des concepts de voisins de Kate dans le contexte de la famille royale. Chaque pré-concept est représenté par son extension, partagée entre son extension propre et le reste. Les lettres représentent les 7 personnes par leur initiale (sauf A pour Charlotte). Les lettres minuscules dans les arcs représentent l'abstraction de ces personnes par des variables. Le pattern de l'intension d'un pré-concept est l'ensemble des arcs sur le chemin menant à ce pré-concept. Les concepts de voisins, ceux qui ne peuvent être par-

titionnés davantage, sont encadrés et la taille de leur extension est affichée comme distance numérique. On trouve 5 concepts de voisins en plus du concept zéro K. Par ordre de distance croissante, on a le concept D/K des mères avec Diana, le concept W/K des parents d'un garçon et d'une fille avec William, le concepts A/DK des personnes de sexe féminin avec Charlotte en plus de Diana, le concept C/DWK des parents avec Charles en plus de Diana et William, et enfin le concept $GH/CDWKA$ de ceux qui ne partagent rien avec Kate.

Comparé à l'algorithme naïf, cet algorithme peut être incomplet au sens où certains concepts de voisins peuvent ne pas être produits, conduisant à placer certains objets à une plus grande distance conceptuelle qu'ils ne le sont réellement. En effet, les patterns générés P sont des sous-ensemble de I, ce qui contraint les homomorphismes de P vers I à être injectifs, c'est-à-dire que deux variables de P ne peuvent pas se projeter sur un même objet. Par exemple, si $Q(u) = (u, \{p(u, u_1), a(u_1), b(u_1)\})$, alors on va générer les patterns $P_{ab} = \{p(u, u_1), a(u_1), b(u_1)\}$, $P_a = \{p(u, u_1), a(u_1)\}$ et $P_b = \{p(u, u_1), b(u_1)\}$, mais pas le pattern $P_* = \{p(u, u'_1), a(u'_1), p(u, u''_1), b(u''_1)\}$. Si des objets contiennent P_* mais pas P_{ab}, alors ils seront placés dans l'extension propre de P_a ou P_b. Rétablir la complétude suppose de permettre la duplication de variable dans un pattern, ce qui accroît considérablement l'espace de recherche. La forme actuelle de l'algorithme a aussi l'avantage de faciliter l'interprétation des intensions de concepts car ce sont des sous-ensembles de la description de l'objet u.

L'énumération par partitionnement offre plusieurs avantages importants. Tout d'abord, grâce au partitionnement des extensions propres, le nombre de pré-concepts est borné à tout moment par le nombre d'objets, alors même que le nombre de patterns possibles est exponentiel avec le nombre d'arcs. Ensuite, l'algorithme offre beaucoup de flexibilité. Les pré-concepts peuvent être partitionnés dans n'importe quelle ordre, autorisant des stratégie en profondeur d'abord ou en largeur d'abord ou l'emploi d'une heuristique. Le choix de l'hyperarc est également libre et indépendant d'un pré-concept à l'autre. Des optimisations peuvent être appliquées telles que : éliminer u de l'extension propre initiale, stopper le partitionnement sur les pré-concepts dont l'extension propre est un singleton. De plus, on peut tout de même utiliser cet algorithme pour calculer la distance conceptuelle avec un objet donné v, simplement en posant $C_{init}.proper = \{v\}$. Enfin, et c'est peut-être le plus important, l'algorithme est *anytime* puisqu'une partition des objets en concepts est définie à tout moment. Cette partition est simplement moins fine si on arrête le processus avant la fin. Pour un parcours en largeur, cela permet de limiter la profondeur d'exploration du graphe par un temps de calcul plutôt que par une profondeur limite.

5.2 Jointure "paresseuse"

Le calcul explicite de l'ensemble $C.homs$ des homomorphismes d'un pattern par jointures successives peut être impratiquable, même dans des cas simples. Par exemple, le pattern $\{film(u), acteur(u, u_1), ..., acteur(u, u_n)\}$ aura de l'ordre de Nn^n homomorphismes, en considérant qu'il y a N films, chacun relié à n acteurs. Pour 1000 films reliés chacun à 10 acteurs, cela fait déjà 10^{13} homomorphismes ! Ce nombre peut diminuer quand des contraintes sont ajoutées aux acteurs, mais peut aussi augmenter de façon exponentielle avec l'ajout de relations, par exemple des acteurs vers leurs films. Il est possible de faire mieux car ce qui nous intéresse *in fine* est $C.ext \subseteq O$, qui est de taille bornée par le nombre d'objets. L'idée est de représenter le homset $C.homs$ par une structure contenant plusieurs jointures locales au lieu d'une jointure globale, en joignant juste ce qu'il faut pour que $C.ext$ soit correct par rapport à $C.pattern$.

Définition 7 (arbre d'homsets)
Un homset factorisé *est une structure arborescente* $\psi = (e, D, \Phi, \Delta, \Psi)$ *où :*
— e *est un hyperarc et* $vars(e)$ *est l'ensemble de ses nœuds-variables ;*
— $D \subseteq vars(e)$ *est l'ensemble des variables introduites par cet hyperarc ;*
— Φ *est un homset dont le domaine contient* $vars(e)$ *;*
— $\Delta \subseteq dom(\Phi)$ *est le sous-domaine utile aux structures englobantes ;*
— Ψ *est l'ensemble des sous-structures arborescentes (structures filles).*

Le homset explicite est égal à la jointure des homsets de l'ensemble des (sous-)structures de l'arbre d'homsets. L'arbre d'homsets du pré-concept initial correspond à l'arc neutre $\top(u)$.

L'*arbre d'homsets initial* correspond au pré-concept initial C_{init} et est défini comme $\psi_{init} := (\top(u), \{u\}, homs_K(\{\top(u)\}), \{u\}, \emptyset)$. Lors de la partition d'un pré-concept, la jointure d'homsets $C.homs \bowtie homs(\{e^*\})$ doit être remplacée par l'évaluation de $lazyjoin(\psi, \psi^*)$ (voir Algorithm 3), avec $\psi = C.homs$ et $\psi^* = (e^*, vars(e^*) \setminus vars(C.pattern), homs_K(\{e^*\}), vars(e^*) \cap vars(C.pattern), \emptyset)$. Cet jointure "paresseuse" consiste à insérer la structure feuille ψ^* correspondant au nouvel arc dans l'arbre d'homsets ψ.

Algorithme 3 (Définition récursive de la fonction $lazyjoin(\psi, \psi^*)$**)**
Require: *deux homset factorisés* $\psi = (e, D, \Phi, \Delta, \Psi)$ *et* $\psi^* = (e^*, D^*, \Phi^*, \Delta^*, \Psi^*)$
Ensure: *deux ensembles* $\Delta^+, \Delta^- \subseteq \mathcal{V}$*, un nouvel homset factorisé* $\psi' = (e, D, \Phi', \Delta', \Psi')$
1: $\Delta^+ \leftarrow \emptyset$; $\Delta^- \leftarrow \emptyset$; $\Phi' \leftarrow \Phi$; $\Psi' \leftarrow \emptyset$; $inserted \leftarrow false$
2: ***for all*** $\psi_c \in \Psi$ ***do***
3: $\Delta_c^+, \Delta_c^-, \psi_c' \leftarrow lazyjoin(\psi_c, e^*)$ *where* $\psi_c' = (e_c, D_c, \Phi_c', \Delta_c', \Psi_c')$
4: $\Delta^+ \leftarrow \Delta^+ \cup \Delta_c^+$; $\Delta^- \leftarrow \Delta^- \cup \Delta_c^-$; $\Phi' \leftarrow \Phi' \bowtie \pi_{\Delta_c'} \Phi_c'$; $\Psi' \leftarrow \Psi' \cup \{\psi_c'\}$
5: ***end for***
6: ***if*** $D \cap \Delta^* \neq \emptyset$ ***then***
7: ***if*** $\neg inserted$ ***then***
8: $\Delta^- \leftarrow \Delta^- \cup (\Delta^* \setminus D)$; $\Phi' \leftarrow \Phi' \bowtie \pi_{\Delta^*} \Phi^*$; $\Psi' \leftarrow \Psi' \cup \{\psi^*\}$; $inserted \leftarrow true$
9: ***else***
10: $\Delta^+ \leftarrow \Delta^+ \cup (\Delta^* \cap D)$
11: ***end if***
12: ***end if***
13: $\Delta^+, \Delta^- \leftarrow \Delta^+ \setminus \Delta^-,\ \Delta^- \setminus \Delta^+$
14: $\Delta' \leftarrow \Delta' \cup \Delta^+ \cup \Delta^-$
15: ***return*** $\Delta^+, \Delta^-, \psi'$

La figure 3 donne un exemple de jointure paresseuse partant de l'arbre d'homsets du PGP $(k, \{female(k), parent(g, k), male(g), parent(g, w), parent(a, k), female(a)\})$ et ajoutant l'arc $e^* = parent(a, w)$ qui forme un cycle en joignant a et w. Les modifications sont propagées à partir des deux sous-structures définissant (voir D) les variables a, w et la nouvelle sous-structure ψ^* est insérée dans l'une d'elle (ici, celle définissant a). Les insertions des autres arcs, menant au premier arbre de la figure, ne modifient qu'une branche de l'arbre à la fois car elles n'introduisent pas de cycle. Dans l'algorithme 3, le calcul des Δ^+, Δ^- permet de déterminer jusqu'où propager dans les Δ les variables définies dans les autres branches que la branche d'insertion (ici, la variable w) pour correctement prendre en compte le cycle dans le calcul des homsets.

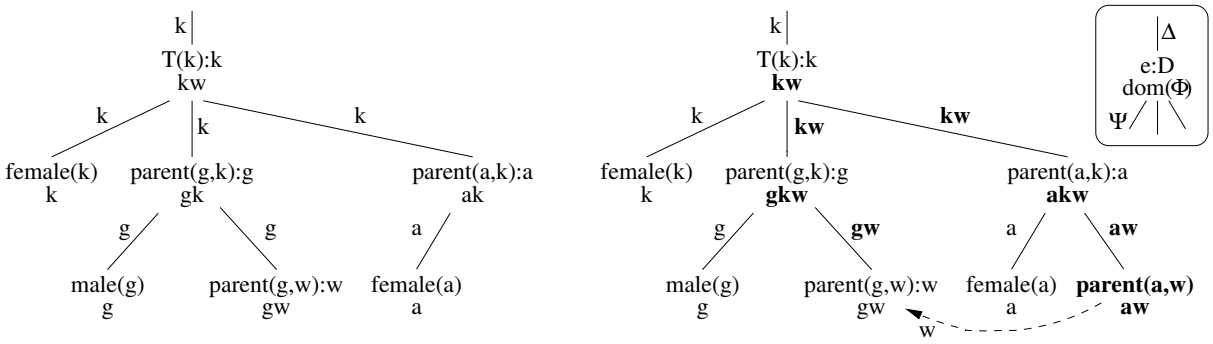


FIGURE 3 – Arbre d'homsets avant et après jointure paresseuse avec l'arc $parent(a, w)$.

Si on reprend l'exemple des films reliés à leurs acteurs, on constate que chaque nouvel arc $e^* = acteur(u, u_i)$ n'entraîne que la projection $\Phi_{e^*} = \pi_{\{u\}} homs_K(\{e^*\})$, c'est-à-dire le calcul du domaine d'une relation binaire (ici, les films ayant un acteur), puis la jointure $\Phi_{\top(u)} \bowtie \Phi_{e^*}$, c'est-à-dire une intersection de deux ensembles d'objets (ici, des films). L'arbre d'homsets final contient donc un homset de cardinal N (arc $\top(u)$) et n homsets de cardinal Nn (arcs $acteur(u, u_i)$), soit un total de l'ordre de Nn^2 homomorphismes au lieu de Nn^n. Pour 1000 films reliés à 10 acteurs chacun, cela fait 10^5 au lieu de 10^{13} ! De plus, les homomorphismes portent sur 1 ou 2 variables au lieu de $(n + 1)$.

6 Implémentation et premières expérimentations

Nous avons implémenté les algorithmes ci-dessus en ∼700 lignes de code OCaml et les avons intégré dans SEWELIS [2] comme amélioration d'UTILIS (Hermann *et al.*, 2012) pour la saisie guidée de descriptions RDF. UTILIS est aussi une méthode de plus proches voisins, mais basé sur une distance numérique et ne traitant ni les cycles, ni l'exploration profonde du graphe de connaissances. Les graphes de connaissances sont ici des graphes RDF. Dans notre implémentation, nous avons adopté les heuristiques suivantes : exploration en largeur d'abord de l'arbre de partitionnement des concepts de voisins ; choix de l'arc de partitionnement favorisant à la fois les arcs les plus proches de l'objet requête u et la diversité des prédicats dans les patterns. Le dernier critère permet d'éviter, par exemple, de considérer 10 acteurs d'un film avant même d'avoir considéré d'autres critères tels que le directeur ou la date de sortie.

Nous avons conduit de premières expérimentations de l'approche CNN sur deux versions de la base MONDIAL (May, 1999), qui contient des données géographiques (ex., pays, continents, rivières, montagnes, langues, groupes ethniques). La version *complète* est un contexte graphe de 41577 nœuds et 120546 arcs. La version *réduite* en est un sous-graphe de 9692 nœuds et 11691 arcs, soit avec environ 5 fois moins de nœuds et 10 fois moins d'arcs. Elle a été obtenue en éliminant les données numériques car elles sont pour l'instant traitées comme des données catégorielles et ne permettent donc guère de trouver des points communs entre objets [3]. La figure 4 montre pour les deux versions de MONDIAL l'évolution de deux mesures selon le temps de calcul alloué (*timeout*), allant de 1s à 200s. La première mesure (graphique gauche) est le

2. `http://www.irisa.fr/LIS/softwares/sewelis/`

3. Par exemple, il y a peu de chances que deux pays aient exactement la même population.

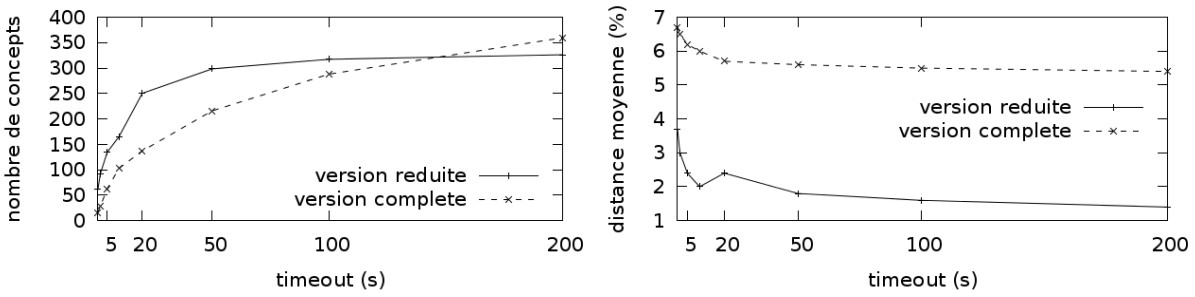


FIGURE 4 – Évolution du nombre de concepts de voisins (à gauche) et de la distance numérique moyenne des objets (à droite) en fonction du *timeout* et de la version de MONDIAL.

nombre de concepts de voisins obtenus. La deuxième mesure (graphique droit) est la distance numérique $d(u, v) = |\delta(u, v).ext|$ moyenne entre l'objet requête u et les autres objets v. Cette dernière est la moyenne des tailles des extensions des concepts de voisins, pondérée par le nombre d'objets dans l'extension propre de chaque concept. La distance numérique est donnée en pourcentage de la distance maximale possible, à savoir le nombre d'objets du jeu de données. Ces mesures sont moyennées sur un échantillon de 100 objets pour les timeouts inférieurs à 20 et sur 10 objets au-delà. Nous avons pu vérifié que ces mesures sont peu sensibles à la taille de cet échantillon : $\pm 30\%$ au maximum en faisant varier cette taille de 1 à 100.

La forme asymptotique des courbes de la figure 4 montre la validité de l'utilisation *anytime* de notre algorithme puisque la plupart des concepts sont produits rapidement et que la distance moyenne décroît d'abord rapidement puis très lentement. Le nombre de concepts est assez grand pour être discriminant et en même temps assez petit comparé au nombre d'objets pour démontrer une capacité d'abstraction. Nous avons aussi regardé le nombre de *plus proches* concepts : il évolue de 2 à 4 pour les deux versions. Les distances moyennes peuvent paraître faibles, mais elles correspondent tout de même à des centaines d'objets dans la version réduite et à des milliers d'objets dans la version complète. Toutes ces mesures montrent que l'on a pas d'écrasement du spectre des distances, ni vers le bas ni vers le haut. Enfin, nous avons évalué l'impact de nos deux optimisations en les désactivant. Si on utilise des jointures explicites plutôt que paresseuses, on produit environ 5 fois moins de concepts avec une distance moyenne environ 3 fois supérieure pour des timeouts de 10s et 100s. Si on applique notre algorithme à chaque objet v isolément ($C_{init}.proper = \{v\}$), en partageant un timeout de 100s entre les 9692 objets de la version réduite, on observe qu'on ne parvient à calculer une distance conceptuelle non-triviale (i.e., pattern non vide) que pour 16% des objets, et que la distance moyenne est 8 fois supérieure.

Pour illustrer notre approche, nous donnons ci-dessous quelques-uns des concepts de voisins les plus proches du pays France dans la version réduite, avec leurs extensions (même notation qu'en figure 2) et leurs intensions (informellement par souci de clarté et de concision) :

— Pologne : une république européenne, voisine de l'Allemagne ;
— Italie : une république européenne où on parle français ;
— Espagne : un pays européen sur l'Atlantique, ayant un voisin et des dépendances (ex., Melilla)
— Royaume-Uni / Espagne : un pays européen sur l'Atlantique avec des dépendences (ex., Cayman).
— Portugal / Espagne : un pays européen sur l'Atlantique, ayant un voisin et une ancienne colonie (ex., Cap vert).

7 Conclusion et perspectives

Nous avons introduit la notion de "concepts de voisins" comme forme conceptuelle de distance entre objets, où ces objets sont les nœuds d'un graphe de connaissances. Nous avons proposé un algorithme pour les calculer et montré qu'il était assez efficace pour être utilisé sur des données réelles et de taille conséquente. Ces résultats sont encore préliminaires et il reste surtout à explorer comment ces concepts de voisins peuvent être exploités au mieux pour des tâches d'apprentissage, supervisé ou non. Il reste des marges d'optimisation de l'algorithme et il sera nécessaire de mieux traiter les valeurs telles que nombres, dates ou chaines.

Références

BISSON G. (2000). La similarité : une notion symbolique/numérique. *Apprentissage symbolique-numérique*, **2**, 169–201.

CHEIN M. & MUGNIER M.-L. (2008). *Graph-based knowledge representation : computational foundations of conceptual graphs*. Advanced Information and Knowledge Processing. Springer.

DE MANTARAS R. L., MCSHERRY D., BRIDGE D., LEAKE D., SMYTH B., CRAW S., FALTINGS B., MAHER M. L., T. COX M., FORBUS K. *et al.* (2005). Retrieval, reuse, revision and retention in case-based reasoning. *The Knowledge Engineering Review*, **20**(03), 215–240.

FERRÉ S. (2015). A proposal for extending formal concept analysis to knowledge graphs. In J. BAIXERIES, C. SACAREA & M. OJEDA-ACIEGO, Eds., *Int. Conf. Formal Concept Analysis (ICFCA)*, LNCS 9113, p. 271–286 : Springer.

FERRÉ S. & CELLIER P. (2016). Graph-FCA in practice. In O. HAEMMERLÉ, G. STAPLETON & C. FARON-ZUCKER, Eds., *Int. Conf. Conceptual Structures (ICCS) - Graph-Based Representation and Reasoning*, LNCS 9717, p. 107–121 : Springer.

FERRÉ S. & RIDOUX O. (2002). The use of associative concepts in the incremental building of a logical context. In G. A. U. PRISS, D. CORBETT, Ed., *Int. Conf. Conceptual Structures*, LNCS 2393, p. 299–313 : Springer.

GANTER B. & WILLE R. (1999). *Formal Concept Analysis — Mathematical Foundations*. Springer.

HAHN G. & TARDIF C. (1997). Graph homomorphisms : structure and symmetry. In *Graph symmetry*, p. 107–166. Springer.

HERMANN A., FERRÉ S. & DUCASSÉ M. (2012). An interactive guidance process supporting consistent updates of RDFS graphs. In A. TEN TEIJE ET AL., Ed., *Int. Conf. Knowledge Engineering and Knowledge Management (EKAW)*, LNAI 7603, p. 185–199 : Springer.

HORVÁTH T., WROBEL S. & BOHNEBECK U. (2001). Relational instance-based learning with lists and terms. *Machine Learning*, **43**(1-2), 53–80.

KUZNETSOV S. (2013). Fitting pattern structures to knowledge discovery in big data. In P. CELLIER, F. DISTEL & B. GANTER, Eds., *Int. Conf. Formal Concept Analysis*, LNAI 7880, p. 254–266. Springer.

MAY W. (1999). *Information Extraction and Integration with* FLORID *: The* MONDIAL *Case Study*. Rapport interne 131, Universität Freiburg, Institut für Informatik. Available from `http://dbis.informatik.uni-goettingen.de/Mondial`.

MITCHELL T. (1997). *Machine Learning*. McGraw-Hill.

MUGGLETON S. (1995). Inverse entailment and progol. *New Generation Computation*, **13**, 245–286.

ROUANE-HACENE M., HUCHARD M., NAPOLI A. & VALTCHEV P. (2013). Relational concept analysis : mining concept lattices from multi-relational data. *Annals of Mathematics and Artificial Intelligence*, **67**(1), 81–108.

Graphe de connaissances et folksonomie : leur performance comparative dans le calcul de l'affinité

Chun Lu[1, 2], Philippe Laublet[1], Milan Stankovic[1, 2] et Filip Radulovic[2]

[1] Laboratoire STIH, Université Paris-Sorbonne, 28 rue Serpente, 75006 Paris
philippe-laublet@paris-sorbonne.fr
[2] Sépage, 27 rue du Chemin Vert, 75011 Paris
{chun, milstan, filip}@sepage.fr

Résumé : L'affinité est un élément essentiel dans bien des systèmes d'information centrés sur l'utilisateur comme les systèmes de recommandation. Le graphe de connaissances et la folksonomie sont respectivement des jalons importants pour le Web Sémantique et le Web Social. Nonobstant leur trait collaboratif partagé (du moins quelques grands graphes de connaissances le sont), les données codées diffèrent tant par la nature (fait versus expérience) que par la structure (formelle versus lâche). Dans ce papier, nous tentons d'éclaircir leur performance comparative dans la tâche du calcul de l'affinité à travers deux expériences dans le domaine du e-tourisme. Nos résultats montrent que le graphe de connaissances permet de calculer l'affinité avec plus de précision alors que la folksonomie augmente la diversité et la nouveauté. Ces constatations nous ont motivés à développer le Framework d'Affinité Sémantique pour bénéficier de leurs avantages respectifs. L'original de ce papier est publié par ESWC 2017.

Mots-clés : Affinité, recommandation, collaboratif, similarité, sémantique, graphe de connaissances, folksonomie, e-tourisme

1 Introduction

L'affinité entre un utilisateur et une entité (ex. film, musique, artiste) est la probabilité que l'utilisateur soit attiré par l'entité ou réalise une action liée à elle (ex. cliquer, acheter, aimer, partager). L'affinité est un élément essentiel dans bien des systèmes d'information centrés sur l'utilisateur comme les systèmes de recommandation, la publicité en ligne, la recherche exploratoire etc. Parmi les techniques de calcul de l'affinité, celles basées sur le contenu posent l'hypothèse que l'utilisateur aurait une affinité plus élevée avec les entités similaires à celles qu'il a appréciées dans le passé. Le graphe de connaissances et la folksonomie sont respectivement des jalons importants pour le Web Sémantique et le Web Social. Ils ont tous deux boosté les techniques basées sur le contenu grâce au grand nombre de données disponibles sur les entités. Sur le Web Sémantique, les utilisateurs contribuent à la création des graphes de connaissances universels comme DBpedia et Wikidata. Sur le Web Social, les utilisateurs annotent et catégorisent les entités avec des étiquettes libres formant des folksonomies. Nonobstant leur trait collaboratif partagé, les données codées diffèrent tant par la nature que par la structure. Les graphes de connaissances structurent des données factuelles avec une ontologie. Les folksonomies contiennent des données d'expérience avec une structure lâche. Nous donnons un exemple pour illustrer leur différence. Sur DBpedia, le film *dbr:Jumanji* est associé aux faits comme *dbr:Joe_Johnston* par la propriété *dbo:director*, *dbr:Robin_Williams* par *dbo:starring*. Dans la folksonomie de MovieLens[1], le même film est abondamment annoté avec des étiquettes comme « nostagic », « not funny », « natural disaster » etc. Ces étiquettes reflètent l'expérience qu'ont eue les différents utilisateurs et ainsi une sorte d'intersubjectivité qui n'est pas présente dans les graphes de connaissances.

[1] https://movielens.org/

Après une étude approfondie de la littérature (section 2), nous n'avons pas réussi à trouver des éclairages utiles sur l'efficacité comparative de ces deux espaces de données dans la tâche du calcul de l'affinité. Ces deux espaces de données continuant de proliférer sur le web, il est plus que jamais nécessaire de faire la lumière sur cette question. Nous étudions cette question à travers deux expériences dans le domaine du e-tourisme. La première expérience hors ligne est décrite dans la section 3 dont les constatations ont motivé le développement du Framework d'Affinité Sémantique (section 4). La section 5 présente la deuxième expérience qualitative et évalue le framework proposé. La section 6 conclut le papier.

2 Travaux connexes

Depuis plus d'une décennie, les chercheurs étudient de près les liens entre le Web Sémantique et le Web Social. L'idée générale derrière ces efforts est d'augmenter la sémantique du Web Social à l'aide des technologies du Web Sémantique (Bontcheva et Rout, 2014). (Passant et Laublet, 2008) proposent l'ontologie *MOAT* et un framework collaboratif pour guider les utilisateurs à fournir la sémantique des tags (étiquettes) lors du processus d'annotation. (Mika, 2007) propose de construire des ontologies légères à partir des folksonomies. (Cantador et al., 2011) présentent une méthode utilisant le Web Sémantique et le traitement automatique des langues pour classer les en fonction de l'intention derrière l'application des tags. Certains auteurs essaient d'extraire des préférences utilisateurs à partir des folksonomies (Orlandi et al., 2012). D'autres essaient de prouver l'avantage de les utiliser dans les systèmes de recommandation (Semeraro et al., 2012). Du côté des graphes de connaissances, les auteurs les exploitent pour calculer la similarité sémantique entre les entités et les incorporent dans les systèmes de recommandation. Dans (Passant, 2010) et (Piao et Breslin, 2016), les auteurs présentent respectivement *Linked Data Semantic Distance* et *Resource Similarity* qui exploitent DBpedia pour recommander des artistes musicaux. Des variantes de *Spreading Activation* sont utilisées dans des systèmes de recommandation inter-domaines (Kaminskas et al., 2014) et de recherche exploratoire (Marie, 2014).

3 Première expérience

Nous avons mené une première expérience dans un scénario de recommandation des destinations de voyage.

3.1 Jeu de données

Nous avons adapté le jeu de données Yahoo! Flickr Creative Commons 100[2] (YFCC100M) (Thomee et al., 2016). YFCC100M contient 100 millions de photos et de vidéos géo-localisées et datées publiées sur Flickr. Les traitements principaux sont les suivants : 1. Retenir seulement les photos/vidéos avec la meilleure précision de géolocalisation 2. Associer chaque photo/vidéo à un lieu d'intérêt dans un graphe de connaissances (Lu et al., 2016) 3. Trier par ordre chronologique. A l'issue de ces traitements, nous avons obtenu, pour chaque utilisateur, une séquence de voyages contenant les villes qu'il a visitées successivement. Le jeu de données final[3] utilisé dans cette expérience contient 3878 utilisateurs et 705 villes. Les séquences de voyages contiennent en moyenne 5,27 villes.

[2] http://webscope.sandbox.yahoo.com/

[3] Le jeu de données ainsi que certaines autres ressources mentionnées se trouvent à : https://bitbucket.org/sepage/semantic-affinity-framework

3.2 Traitements de la folksonomie et du graphe de connaissances

Nous utilisons une folksonomie collectée sur un site de voyage collaboratif. Elle contient 234 étiquettes sur 26,237 villes dans 154 pays. Nous l'avons modélisée dans un espace vectoriel de type *tag genome* (Vig et al., 2012) où les villes sont notées sur une échelle continue de 0 à 1 pour chaque étiquette. La mesure cosinus est utilisée pour calculer la similarité entre les villes.

Pour chacune des 705 villes dans le jeu de données, nous avons exécuté des requêtes SPARQL avec toutes les propriétés sélectionnées (TABLE 1). Les ressources liées par la propriété *skos:broader* sont assimilées à celles liées par *dct:subject*. Nous avons ensuite éliminé les ressources qui sont liées à une seule ville car elles ne contribuent guère au calcul de la similarité. 501,365 ressources sont initialement obtenues et 29,743 d'entre elles sont finalement retenues. Nous avons adopté la mesure de Jaccard dont l'efficacité a été démontrée dans une comparaison avec des mesures plus sophistiquées (*VsmSim, GbkSim, FuzzySim)* dans un scénario de recommandation musicale (Nguyen et al., 2015).

TABLE 1 – *Propriétés DBpedia sélectionnées pour calculer la similarité entre les villes*

Entrant		**Sortant**	
dbo:birthPlace	dbo:broadcastArea	dbo:isPartOf	dbo:part
dbo:location	dbo:nearestCity	dbo:country	dbo:twinTown
dbo:deathPlace	dbo:ground	dbo:timeZone	dbo:saint
dbo:city	dbo:foundationPlace	dbo:Mayor	dbo:district
dbo:capital	dbo:assembly	dbo:region	dct:subject
dbo:hometown	dbo:restingPlace	dbo:province	(skos:broader)
dbo:recordedIn	dbo:place	dbo:leaderName	
dbo:residence	dbo:locationCity		
dbo:headquarter			

3.3 Calcul de l'affinité

Etant donné un profil utilisateur contenant une liste de villes visitées dans le passé, le score d'affinité d'une ville candidate v_i est la moyenne des scores de similarité qu'elle a avec chacune des villes de son profil.

$$affinité(u, v_i) = \frac{\sum_{v_j \in profil(u)} sim(v_i, v_j)}{|profil(u)|} \qquad (1)$$

3.4 Protocole et métriques

Pour une séquence de voyages de *n* villes, les *n-1* premières villes constituent le profil utilisateur, la *n-ième* ville est considérée comme la vérité terrain. Chaque approche prend le profil utilisateur en entrée et génère trois listes de recommandations contenant respectivement 10, 20, 30 villes dans l'ordre décroissant des scores d'affinité. Dans le scénario de la recommandation, la capacité du calcul de l'affinité est reflétée par la précision. Nous utilisons deux métriques pour la mesurer : Succès (2) et Mean Reciprocal Rank (MRR) (3). Depuis quelques années, les chercheurs qui travaillent sur les systèmes de recommandation portent un intérêt particulier sur la diversité et la nouveauté. Ces deux qualités sont mesurées avec les formules (5) et (6). Par analogie avec (Di Noia et al., 2014), la diversité intra-liste (ILS) est calculée par rapport à deux propriétés : *dbo:country* et *dct:subject*. Nous utilisons les valeurs pagerank des ressources DBpedia comme les scores de popularité. Comme dans (Nguyen et al., 2015), nous considérons que 20% des villes ayant les meilleurs scores sont populaires.

$$Succès = \frac{\sum_{u \in U} rel_{g,u}}{|U|} \text{ où } rel_{g,u} = \begin{cases} 1, si\ vérité\ terrain\ g\ est\ dans\ top - N \\ 0,\ sinon \end{cases} \quad (2)$$

$$MRR = \frac{1}{|U|} \sum_{u \in U} \frac{1}{rank_u} \quad (3)$$

$$ILS_u@N = \sum_{i \in L_u^N} \sum_{j \in L_u^N} \frac{sim(i,j)}{|paires|} \quad (4) \qquad ILS@N = \frac{1}{|U|} \sum_{u \in U} ILS_u@N \quad (5)$$

$$Nouveauté@N = \frac{nombre\ de\ villes\ recommandées\ impopulaires}{N * |U|} \quad (6)$$

3.5 Résultats et discussions

Les résultats sont présentés dans la TABLE 2. Les tests t appariés montrent que les différences entre les deux approches sont statistiquement significatives sur toutes les métriques dans toutes les configurations. Nous observons un avantage net de GC sur FOLK en termes de succès et de MRR, ce qui reflète la capacité de détecter les villes qui sont en affinité élevée avec l'utilisateur et de mieux les ordonner. Les recommandations générées par FOLK sont plus diverses et plus nouvelles. Dans la folksonomie, certains aspects ne sont pas pas présents, tels que la géographie (*dbo:country, dbo:region)*, les personnes (*dbo:birthPlace, dbo:residence*), les catégories connexes (*dct:subject, skos:broader*). La folksonomie contient des traits tels que « Luxury Brand Shopping », « Clean Air » et « Traditional food ». Ces traits peuvent être partagés par de différentes villes du monde même si elles sont moins populaires. Ces constatations nous ont motivés pour développer le Framework d'Affinité Sémantique qui profite de la complémentarité des deux approches pour parvenir à un compromis équilibré sur la pertinence, la diversité et la nouveauté.

TABLE 2 – *Résultats de la première expérience, GC : Graphe de connaissance, FOLK : Folksonomie*

	Top-10		Top-20		Top-30	
	GC	FOLK	GC	FOLK	GC	FOLK
Succès	**0.232**	0.06	**0.33**	0.116	**0.386**	0.166
MRR	**0.047**	0.003	**0.047**	0.003	**0.047**	0.003
ILS	0.257	**0.089**	0.208	**0.072**	0.176	**0.065**
Nouveauté	0.717	**0.824**	0.722	**0.772**	0.723	**0.755**

4 Framework d'Affinité Sémantique

Nous proposons un Framework d'Affinité Sémantique (FAS) qui intègre, agrège, enrichit et nettoie les données sur les entités en provenance des graphes de connaissances et des folksonomies. Son pipeline est décrit dans FIGURE 1. Une explication plus détaillée se trouve dans (Lu et al., 2017). Un graphe d'affinité est généré à l'issue du processus. Nous avons également développé un mécanisme pour expliquer les recommandations. Par exemple, on peut expliquer la recommandation de *dbr:Ljubljana* par *dbc:Capitals_in_Europe*. Etant donné une liste d'entités, nous cherchons les caractéristiques les plus fréquentes tout en contrôlant leur diversité par le biais des propriétés qui relient les caractéristiques aux entités.

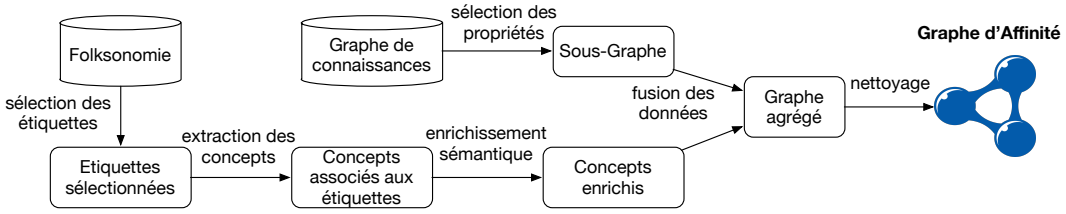


FIGURE 1 – Pipeline du Framework d'Affinité Sémantique

5 Deuxième expérience

Nous avons mené une deuxième expérience qualitative pour évaluer l'utilité et l'efficacité du Framework d'Affinité Sémantique. Nous avons généré un graphe d'affinité avec les données décrites dans 3.2 que nous nommons GA. A part la recommandation, nous nous sommes intéressés à la capacité d'explication des différentes approches.

You submitted:	You might like:	We recommend you:
dbr:Rome	dbc:Clothing	dbr:The_Hague
dbr:Florence	dbr:Food	dbr:Haarlem
dbr:Amsterdam	dbr:David_de_Haen	dbr:Naples
	dbr:Italy	dbr:Milan
	dbr:History	dbr:Turin

FIGURE 2 – Exemple de recommandations et d'explications générées par GA

37 personnes âgées de 25 à 38 ans ont participé à cette expérience dont 19 hommes et 18 femmes. Ils travaillent tous dans des sociétés sises dans une pépinière d'entreprise à Paris. Nous avons demandé aux participants de se mettre dans le scénario de rechercher pour la prochaine destination de voyage. Ils sont allés sur l'interface de test où les 705 villes étaient affichées dans un ordre aléatoire. Ils ont choisi plusieurs villes qui leur paraissaient intéressantes à première vue. Une fois que ces choix ont été soumis, ils ont reçu trois listes de recommandations contenant cinq villes accompagnées d'une explication avec cinq entités/étiquettes/catégories. FIGURE 2 montre un exemple généré par GA. Ils ont noté la pertinence, la diversité et la nouveauté/intérêt des recommandations et des explications sur une échelle de 1 à 5. Nous considérons que les notes supérieures à 3 comme notes positives. Nous utilisons comme métrique le pourcentage des notes positives.

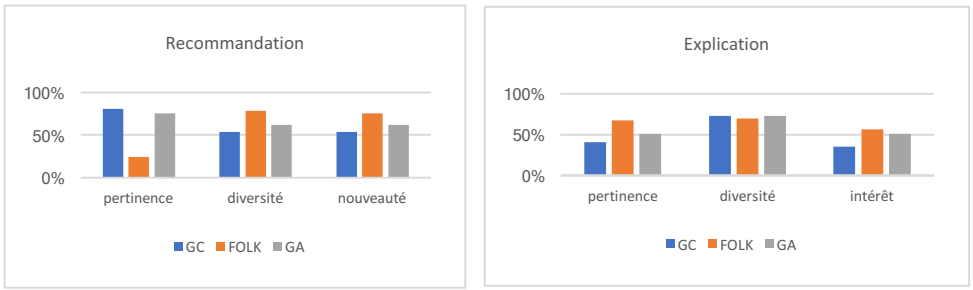


FIGURE 3 – Résultats de la deuxième expérience

Sur la recommandation, les résultats présentés dans FIGURE 3 sont en phase avec ceux de la première expérience. Sur l'explication, les résultats montrent que celle générée par FOLK est la mieux appréciée. Notre folksonomie a été créée de manière collaborative par les voyageurs. Elle couvre plusieurs aspects du voyage (nourriture, activité, transport). La capacité d'explication de GA a été boostée par l'inclusion de la folksonomie, ce qui l'a permis de surperformer GC. Les participants sont en général sceptiques envers GC. Certains trouvent ses explications assez générales, par exemple *dbr:Leisure*. D'autres les trouvent difficiles à comprendre, par exemple *dbr:China_Record_Corporation*. En effet, ces problèmes pourraient être résolues à travers l'utilisation de l'arbre des catégories DBpedia (définir un seuil en deça duquel une catégorie peut être considérée comme étant trop générale) ou le filtrage sur *rdf:type* (mettre certaines classes sur liste noire). Ces filtrages pourraient être intégrés dans le pipeline du Framework d'Affinité Sémantique.

6 Conclusion

Dans ce papier, nous avons étudié la performance comparative du graphe de connaissance et de la folksonomie dans la tâche du calcul de l'affinité à travers deux expériences dans le domaine du e-tourisme. Les résultats montrent que le graphe de connaissances permet de calculer l'affinité avec plus de précision alors que la folksonomie augmente la diversité et la nouveauté. Nous avons développé le Framework d'Affinité Sémantique pour bénéficier de leurs avantages respectifs. La combinaison des deux espaces de données aboutit à une performance équilibrée tant pour la recommandation que pour l'explication. Plus de détails se trouvent dans l'original de ce papier (Lu et al., 2017).

Références

Bontcheva, K., Rout, D. (2014). Making sense of social media streams through semantics: a survey. Semantic Web, 5(5), 373-403

Cantador, I., Konstas, I., Jose, J. M. (2011). Categorising social tags to improve folksonomy-based recommendations. Web Semantics: Science, Services and Agents on the World Wide Web, 9(1), 1-15

Di Noia, T., Cantador, I., & Ostuni, V. C. (2014, May). Linked open data-enabled recommender systems: ESWC 2014 challenge on book recommendation. In *Semantic Web Evaluation Challenge* (pp. 129-143). Springer International Publishing.

Kaminskas, M., Fernández-Tobías, I., Ricci, F., Cantador, I. (2014). Knowledge-based identification of music suited for places of interest. Information Technology & Tourism, 14(1), 73-95

Lu, C., Laublet, P., Stankovic, M. (2016). Travel attractions recommendation with knowledge graphs. In: Proceedings of the 20th International Conference on Knowledge Engineering and Knowledge Management. Bologna, Italy

Lu, C., Stankovic, M., Radulovic, F., & Laublet, P. (2017, May). Crowdsourced Affinity: A Matter of Fact or Experience. In *European Semantic Web Conference* (pp. 554-570). Springer, Cham.

Marie, N. (2014). Linked data based exploratory search. Doctoral dissertation. Université de Nice Sophia-Antipolis

Mika, P. (2007). Ontologies are us: A unified model of social networks and semantics. Web semantics: science, services and agents on the World Wide Web, 5(1), 5-15

Nguyen, P. T., Tomeo, P., Di Noia, T., & Di Sciascio, E. (2015). Content-based recommendations via DBpedia and Freebase: a case study in the music domain. In International Semantic Web Conference (pp. 605-621). Springer International Publishing.

Orlandi, F., Breslin, J., Passant, A. (2012). Aggregated, interoperable and multi-domain user profiles for the Social Web. In: Proceedings of the 8th International Conference on Semantic Systems (pp. 41-48). ACM

Piao, G., and Breslin, J. (2016). Measuring semantic distance for linked open data-enabled recommender systems. In: Proceedings of the 31st Annual ACM Symposium on Applied Computing. ACM

Passant, A.: dbrec—music recommendations using DBpedia. In: Proceedings of the 9th International Semantic Web conference (pp. 209-224). Springer Berlin Heidelberg (2010)

Passant, A., Laublet, P. (2008). Meaning Of A Tag: A collaborative approach to bridge the gap between tagging and Linked Data. In: Proceedings of Linked Data on the Web workshop

Semeraro, G., Lops, P., De Gemmis, M., Musto, C., & Narducci, F. (2012). A folksonomy-based recommender system for personalized access to digital artworks. Journal on Computing and Cultural Heritage (JOCCH), 5(3), 11

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., and Li, L. J. (2016). YFCC100M: The new data in multimedia research. Communications of the ACM, 59(2), (pp. 64-73)

Vig, J., Sen, S., & Riedl, J. (2012). The tag genome: Encoding community knowledge to support novel interaction. ACM Transactions on Interactive Intelligent Systems (TiiS), 2(3), 13.

Modèle de recherche d'information sémantique en graphe : interrogation par propagation d'activation

Ines Bannour, Haïfa Zargayouna, Adeline Nazarenko

LABORATOIRE D'INFORMATIQUE DE PARIS NORD (LIPN, UMR 7030)
Université Paris 13 – Sorbonne Paris Cité & CNRS
Email: prenom.nom@lipn.univ-paris13.fr

Résumé : La recherche d'information sémantique (RIS) cherche à dépasser les approches classiques purement statistiques en injectant des connaissances de manière à désambiguïser le vocabulaire ou enrichir la représentation des documents et des requêtes. Néanmoins, les modèles sémantiques, censés aller au-delà des mots et raisonner à un niveau conceptuel, sont en fait eux-mêmes « aplatis » : on ne représente plus les documents comme des « sacs de mots » mais comme des « sacs de concepts ».

Nous proposons de modéliser les données sémantiques et documentaires sous la forme de graphe pondéré et l'interrogation comme une propagation d'activation dans ce graphe à partir des nœuds qui ont été activés par la requête de l'utilisateur. Cet algorithme a le mérite de préserver les caractéristiques largement éprouvées des modèles classiques de recherche d'information tout en permettant une représentation adéquate des modèles sémantiques. La propagation d'activation sur ce graphe est le mécanisme qui assure la mise en correspondance entre le besoin de l'utilisateur formulé sous la forme d'une requête et les documents. Selon que l'on introduit ou pas de la sémantique dans le graphe, cette approche permet de reproduire une RI classique ou assure en sus certaines fonctionnalités sémantiques. Ces fonctionnalités sont validées expérimentalement sur un corpus dans le domaine médical (Ohsumed87) qui permet de vérifier le passage à l'échelle de notre modèle et de faire une première analyse qualitative et quantitative des performances de l'algorithme de propagation.

Mots-clés : graphe, ontologie, annotation sémantique, interrogation, propagation d'activation

1 Introduction

La recherche d'information (RI) consiste à retrouver des documents classés par ordre de pertinence qui répondent à une requête utilisateur généralement exprimée sous le formes de mots-clés. Les modèles de recherche d'information définissent une représentation des documents et des requêtes ainsi qu'une fonction de correspondance qui permet de calculer des similarités entre documents et requêtes et de classer les résultats. Quel que soit le modèle de recherche d'information et le paradigme sous-jacent (géométrique, probabiliste ou logique), les calculs sont purement numériques et reposent essentiellement sur la fréquence des mots et l'analyse de leur distribution. La recherche d'information sémantique (RIS) cherche à dépasser cette approche formelle, en injectant des connaissances de manière à désambiguïser le vocabulaire ou enrichir la représentation des documents et des requêtes.

Les travaux en recherche d'information sémantique sont nombreux et consistent pour la plupart à adapter les modèles classiques de RI (Zargayouna *et al.*, 2015) mais on observe que les modèles sémantiques, qui sont sensés permettre d'aller au-delà des mots et raisonner à un niveau conceptuel, sont en fait eux mêmes aplatis et on ne fait que passer d'une représentation des documents en « sacs de mots » à une représentation en « sacs de concepts ». Les calculs de correspondance intègrent les calculs de similarité sémantique entre concepts mais les liens sémantiques entre les mots ne sont pas explicités et exploités, alors même que c'est l'un des intérêts principaux des modèles sémantiques. La RIS semble atteindre un plateau, en dépit de l'intérêt toujours croissant qu'elle suscite et de l'influence des nouvelles technologies séman-

tiques du web de données et on peut se demander si l'approche consistant à adapter les modèles documentaires classiques par injection de sémantique n'a pas atteint ses limites.

Nous cherchons à développer un modèle de RIS qui combine au mieux les deux modèles de base : il s'agit d'intégrer un modèle sémantique dans un modèle documentaire sans pour autant perdre la richesse de la structure des ontologies ou des ressources sémantiques utilisées et sans abandonner non plus les calculs distributionnels qui font la force et la robustesse de la RI classique. Nous présentons dans ce travail, un nouveau modèle pour la recherche d'information sémantique qui repose sur une intégration du modèle documentaire et sémantique dans une représentation en graphe, la fonction de correspondance consistant à appliquer une fonction de propagation dans ce graphe.

Nous présentons dans un premier temps un état de l'art sur les travaux qui exploitent les ressources sémantiques en RI et nous dressons le bilan des approches sémantiques à base de graphes (section 2). Dans la section 3, nous détaillons notre modèle et expliquons le mécanisme de propagation dans les graphes que nous instantions. La section 4 présente les expériences faites sur la collection médicale *Ohsumed*.

2 État de l'art

2.1 Approches de RIS

L'exploitation des ressources sémantiques externes a pour but de pallier aux problèmes des modèles classiques de RI, comme le modèle vectoriel ou le modèle probabiliste.

S'attaquer aux problèmes des modèles classiques de type « sac de mots », revient à injecter des connaissances, comme des couples de synonymes ou les concepts du domaine, de manière à désambiguïser le vocabulaire ou enrichir la représentation des documents et des requêtes. L'objectif sous-jacent est de permettre à un système de recherche d'information de renvoyer un document contenant non pas les mots de la requête mais des mots sémantiquement proches et ainsi de réduire le silence, ou, à l'inverse, d'exclure des documents ambigus et de réduire le bruit.

Le but de l'enrichissement de la représentation des documents est d'aboutir à une représentation fidèle des documents (et des requêtes) qui soit plus riche et moins ambiguë qu'une représentation classique par mots. Étant donné l'ambiguïté des ressources sémantiques les plus utilisées en RIS – notamment le thésaurus WordNet et les thésauri médicaux, UMLS et MeSH – leur utilisation en RIS repose souvent sur un processus préalable de désambiguïsation (voir entre autres les travaux de Stetina *et al.* (1998); Guarino *et al.* (1999); Khan (2000); Baziz *et al.* (2003)). Les étapes d'une indexation sémantique conceptuelle, qui tend à enrichir la représentation des documents, sont (i) l'annotation sémantique qui assure le lien entre les unités du texte et les unités sémantiques de la ressource et (ii) le choix des unités d'index et leur calcul de pondération. Ces unités d'index peuvent être des termes, des concepts ou encore une combinaison de termes et de concepts pour pallier le défaut de couverture sémantique de la ressource (Dinh & Tamine, 2010; Hamadan *et al.*, 2012).

L'expansion des requêtes peut être réalisée en étendant le vocabulaire des requêtes au moyen de termes similaires (généralement des synonymes). En 1968, Salton constate déjà que l'utilisation du thésaurus *Harris Synonym* permet d'améliorer les performances, à condition que les termes utilisés pour l'enrichissement soient validés manuellement par un documentaliste.

Il constate également que l'expansion automatique utilisant l'ensemble des termes possibles, dégrade ces performances (Salton, 1968). Plusieurs méthodes d'expansion de requêtes existent dans l'état de l'art : Bhogal *et al.* (2007) passent en revue celles qui exploitent des ontologies.

Zargayouna *et al.* (2015) dressent un état de l'art de ces modèles classiques adaptés pour la RIS. Un bilan est difficile à dresser, l'évaluation d'un système de RIS reste complexe car la qualité des résultats dépend fortement de la manière dont les ressources sont exploitées ainsi que des annotations qui ne sont possibles que si l'on dispose de ressources suffisamment couvrantes.

2.2 Approches à base de graphes

Les travaux en Web Sémantique (WS) adoptent une modélisation en graphe de connaissances. Les documents sont représentés comme des instances, l'accès à ces documents se fait *via* les annotations sémantiques. Cela pose le problème de couverture des ressources, puisque toute information non annotée est perdue. De plus, les interrogations étant exactes, les réponses aux requêtes ne sont pas classées.

Les travaux récents en Web Sémantique proposent plus ou moins d'intégrer des moteurs classiques à des moteurs d'interrogation du WS (Zhang *et al.*, 2007; Fernández *et al.*, 2011; Wang *et al.*, 2011). Narula & Jain (2014) présentent un aperçu de quelques systèmes. Dans les approches de ce type, les documents retrouvés viennent en appui à la réponse factuelle apportée par le système, ils ne constituent pas le cœur de la réponse du système. Ces modèles d'accès sur le WS sont des modèles de recherche de données qui ont été adaptés pour la recherche de documents mais où le document est une information accessoire donnant la provenance des connaissances.

La RI a cependant intégré la notion de graphe bien avant l'avènement du WS. Il s'agissait de prendre en compte la structure de graphe du web documentaire, indépendamment des couches sémantiques pouvant s'y ajouter. Le web est un vaste réseau hypertexte, c'est-à-dire un graphe de documents reliés par des liens de citation encodés sous la forme de liens html orientés. Dans les années 1990, l'idée est apparue que la structure de ce vaste réseau de citation pouvait être utilisée pour estimer la notoriété des pages web et améliorer leur classement dans les résultats des moteurs de recherche. Diverses approches ont été proposées même si c'est l'algorithme PageRank qui est le plus connu (Brin & Page, 1998).

Nous proposons un modèle dédié à la RIS, qui représente les documents sous la forme de graphes intégrant des caractéristiques du modèle documentaire et du modèle sémantique. Ce modèle de RI permet de représenter de manière homogène des informations de natures différentes et de les intégrer dans un même calcul de pertinence. Nous présentons dans ce qui suit notre modèle en graphe, l'algorithme de propagation d'activation et les fonctionnalités sémantiques que ce type de raisonnement permet d'assurer.

3 Modèle en graphe et propagation d'activation

Notre modèle de RIS permet de représenter les données textuelles et leurs propriétés statistiques (fréquences d'occurrences, etc.) ainsi que les connaissances sémantiques issues d'ontologies dans un même *réseau sémantico-documentaire*. Nous intégrons les relations sémantiques des ontologies et les relations termes-documents de la RI traditionnelle dans un unique modèle

de *graphe pondéré* et nous modélisons la fonction de correspondance requête-résultats sous la forme d'un mécanisme de *propagation d'activation* dans le graphe.

3.1 Modélisation en graphe

Nous proposons de représenter la base documentaire et le modèle sémantique qui lui est associé sous la forme d'un réseau sémantico-documentaire. Cette structure permet d'introduire différents types de nœuds et différents types de relations selon ce qu'on souhaite représenter.

Le réseau *sémantico-documentaire* comporte de ce fait trois types de nœuds : nœuds documents (N_d), nœuds termes (N_t) et nœuds concepts (N_c). Ces nœuds sont liés les uns aux autres par cinq types de relations qui peuvent porter certaines propriétés numériques ou symboliques : les relations d'occurrence (R_{occ}), d'intertextualité (R_{int}), terminologiques (R_{ter}), lexicales (R_{lex}) d'annotation (R_{ann}) et ontologiques (R_{ont}) (Bannour *et al.*, 2016).

Différentes configurations du réseau sémantico-documentaire sont possibles, selon les connaissances qu'on choisit de représenter et selon les propriétés symboliques ou numériques qu'on décide d'attribuer aux nœuds et aux arcs du réseau. Ce réseau a vocation à être utilisé pour différentes applications (RI, accès aux données, catégorisation, etc.) mais il faut le paramétrer en orientant et en pondérant les liens qui le composent, ainsi qu'en introduisant les éléments nécessaires au calcul distributionnel. Ces paramétrages sont dépendants des calculs à effectuer sur le graphe. Nous reviendrons sur cet aspect dans la section suivante.

Nous proposons de représenter ce type de réseau sémantique sous la forme d'un *graphe pondéré*, $G = \langle N, R \subseteq N \times \mathcal{R} \times N \rangle$, qui est constitué d'un ensemble de nœuds ($N = N_d \uplus N_t \uplus N_c$) et d'arcs orientés et pondérés ($R = R_{occ} \uplus R_{int} \uplus R_{ter} \uplus R_{lex} \uplus R_{ann} \uplus R_{ont}$).

Les arcs traduisent les relations qu'entretiennent les différents nœuds. Les pondérations peuvent être binaires ou calculées (relations booléennes ou numériques) pour prendre en compte la force des liens exprimés par les relations. Ces valeurs expriment des propriétés intrinsèques de ces liens, des propriétés qu'on ne peut pas retrouver par calcul ou dériver de la structure du graphe, par exemple, mais l'interprétation de ces valeurs dépend des types des nœuds qui sont mis en relation :

le poids d'une relation d'occurrence représente la fréquence d'occurrence d'un terme dans un document ; il peut être normalisé ou non et permet de garder une trace de l'importance des termes dans un document ;

le poids d'une relation terminologique permet de distinguer, par exemple, le label « préféré » d'un concept par rapport aux autres termes qui lui sont associés ;

le poids d'une relation lexicale indique si deux termes sont variantes l'un de l'autre et éventuellement le degré de confiance qu'on a en cette relation ;

le poids d'une relation d'annotation indique si un concept est associé à un document ou non ; ce peut être une valeur booléenne ou une valeur de confiance fournie par l'outil de catégorisation ;

le poids d'une relation d'intertextualité indique si deux documents sont reliés (valeur booléenne) et éventuellement la force de cette relation (valeur numérique) [1] ;

1. Il ne s'agit pas d'une mesure de similarité documentaire, laquelle n'est pas une propriété intrinsèque des documents reliés car elle se calcule sur l'ensemble de la collection.

le poids d'une relation ontologique indique s'il y a une relation hiérarchique ou sémantique entre deux concepts et en donne éventuellement l'importance ; en jouant sur les valeurs de ces liens, on peut activer ou désactiver certains types de liens pour la recherche. On peut ainsi choisir par exemple de privilégier les liens de spécialisation dans la recherche ou au contraire de bloquer les parcours inverses, en leur attribuant des poids nuls ou négatifs, etc. [2].

Les arcs peuvent être orientés ou non, selon le type du lien :

les relations terminologiques, lexicales, d'occurrences et d'annotations peuvent être parcourues dans les deux sens et ne sont pas orientées ;

les relations d'intertextualité peuvent être orientées ou non (par ex. les relations `est_cité_par` et `cite` n'ont pas nécessairement le même poids [3]) selon le type de la relation considérée et les choix de modélisation ;

les relations ontologiques sont généralement considérées comme orientées quand il s'agit de relations hiérarchiques mais certaines relations peuvent aussi être considérées comme symétriques, si c'est explicité dans le modèle sémantique.

La représentation unifiée de toutes ces informations sous la forme d'un graphe pondéré permet d'interroger les documents de manière plus riche que par les seuls mots-clés. On peut accéder au graphe par plusieurs points d'entrée selon que la requête comporte des termes, des documents, des concepts, une combinaison de différents types de nœuds, etc. De même, les réponses attendues peuvent être de différents types. Selon l'application, un filtrage est effectué sur l'ensemble des nœuds retournés pour sélectionner le/les type(s) de nœud(s) qu'on recherche. Le modèle permet ainsi de prendre en compte diverses formes de requêtes et de proposer différents types de résultats, sans avoir à changer de système d'accès à l'information ou de langage d'interrogation.

La fonction de correspondance que nous proposons sur ce graphe pondéré, consiste à appliquer une méthode de *propagation d'activation*, qui part des nœuds de la requête et active de proche en proche les nœuds voisins dans le graphe.

3.2 Appariement par propagation d'activation

La propagation d'activation (PA) est un processus qui permet de propager une information de proche en proche sur un graphe. Elle reprend une idée ancienne issue de la psychologie cognitive et l'intelligence artificielle (Quillian, 1968), selon laquelle les unités lexicales ne sont pas isolées mais prises dans un réseau de relations de sorte que l'activation en mémoire d'une unité active aussi par association les unités voisines.

Dans le cadre de l'accès à l'information dans le WS, la propagation d'activation sur les graphes de connaissances RDF a fait l'objet de plusieurs travaux qui proposent une hybridation de la RI et du WS avec une interrogation par mots-clés (Rocha *et al.*, 2004; Jiang & Tan, 2006; Schumacher *et al.*, 2008). Ces travaux nécessitent le plus souvent d'avoir des ontologies avec

2. La similarité entre concepts dépend plutôt de la distance dans l'ontologie par exemple et n'est donc pas une propriété intrinsèque.

3. Dans le domaine juridique, Mimouni *et al.* (2014) évoquent par exemple la relation de transposition entre une directive européenne et un texte réglementaire ou législatif national.

une composante textuelles riche afin d'éviter le silence et de garantir une amélioration de la recherche par la prise en compte de la sémantique.

Crestani (1997) a formalisé le problème de la propagation d'activation pour la RI et Brouard (2013) a prouvé mathématiquement la correspondance entre son modèle de PA et le modèle vectoriel.

Nous détaillons dans ce qui suit le mécanisme de propagation d'activation que nous appliquons au graphe pondéré.

La propagation d'activation consiste en un ensemble d'*étapes de propagation* et un ou plusieurs *mécanismes de contrôle* de la propagation. A chaque étape, les *valeurs d'activation* associées aux nœuds du graphe sont calculées. Une *étape de propagation* se décompose en deux phases :

l'activation consiste à sélectionner les nœuds à activer parmi les nœuds dont la valeur d'activation est non nulle et qui n'ont pas encore été activés, puis à déclencher la propagation à partir de ces nœuds-là ;

la propagation transmet l'activité d'un ou plusieurs nœuds sources vers leurs voisins avant de désactiver les nœuds sources et de recalculer les *valeurs d'activation* des nœuds cibles.

A une étape de propagation n, l'*activation* s'applique aux nœuds dont les valeurs d'activation ont été mises à jour par « contagion » à partir de leurs voisins directs à l'étape de propagation précédente $(n-1)$. Ce processus est répété en suivant les même phases d'*activation* et de *propagation* jusqu'à ce que plus aucun nœud ne puisse être activé (sélectionné). La propagation d'activation se termine quand aucun nœud ne peut plus être sélectionné et que la distribution des valeurs d'activation sur le graphe s'est stabilisée. Comme les graphes peuvent contenir des cycles, il n'est pas garanti que le processus de propagation itératif se stabilise et il faut introduire un mécanisme de contrôle.

Dans ce travail, nous n'utilisons pas de contraintes spécifiques pour limiter la propagation sur le graphe, ni de contraintes dépendant du domaine et de l'application, comme celles qui sont présentées par Crestani (1997) : contrainte de distance, de chemin, *fan-out* et de seuil. Le mécanisme de contrôle fait partie intégrante de l'algorithme de propagation d'activation et repose sur *la modification de l'état des nœuds du graphe*.

Au départ, l'ensemble des nœuds *actifs* contient les nœuds correspondant à la requête, tandis que les autres nœuds apparaissent comme *inactifs*. Le calcul des *valeurs d'activation* ne s'effectue que sur les nœuds *inactifs* atteints au cours de la propagation. A la fin d'une étape de propagation, les nœuds *actifs* sont *désactivés* tandis que les nœuds *inactifs* dont la valeur est non nulle à l'issue du calcul de propagation constituent les nœuds *actifs* de l'étape de propagation suivante. Un nœud désactivé ne peut plus être réactivé mais sa valeur d'activation peut continuer à croître sous l'influence de ses voisins. La terminaison de l'algorithme est assurée : il s'arrête quand l'ensemble des nœuds *actifs* est vide, c'est-à-dire au plus tard au bout de $|N|$ itérations, où N est le nombre de nœuds du graphe.

A chaque étape de propagation, de nouvelles valeurs d'activation a_k sont calculées sur la base des valeurs d'activation a_{k-1} et en fonction de la structure du graphe. Soient un nœud i et $a_{k-1}(x)$ la valeur d'activation d'un noeud x à l'issue de l'itération $k-1$. La valeur d'activation de i à l'itération k est définie par l'équation 1 suivante :

$$a_k(i) = a_{k-1}(i) + \sum_{j \in pred(i) \cap actif(k-1)} \frac{a_{k-1}(j) * w(j,i)}{deg(j)} \quad (1)$$

où $pred(i)$ retourne la liste des nœuds « prédécesseurs » de i, qui pointent vers le nœud i, $actif(k)$ est l'ensemble des nœuds actifs à l'itération k, $w(j,i)$ est la valeur de l'arc reliant i à j et $deg(j)$ est le degré du nœud j.

Les valeurs d'activation dépendent de *la structure du graphe* et de *l'état des nœuds du graphe*. Le calcul des valeurs d'activation à l'étape k dépend des nœuds prédécesseurs actifs à l'étape $k-1$ (qui sont désactivés à l'étape k) et de leurs valeurs d'activation à l'étape $k-1$. Du fait des cycles, les valeurs d'activation de ces prédécesseurs peuvent être réévaluées ultérieurement mais sans que cela n'affecte leurs voisins. Ceci rend la fonction de mise à jour *synchrone* [4].

3.3 Fonctionnalités sémantiques

L'interrogation du graphe pondéré par différents points d'entrée, permet de :

— poser des requêtes en utilisant des termes qui n'existent pas dans le vocabulaire de la collection, mais peuvent appartenir au volet terminologique de la ressource sémantique, ce qui répond au problème de *term mismatch* décrit par Crestani (2000) ;

— poser des requêtes par des concepts de catégorisation qui ne sont pas forcément associés à un terme du vocabulaire mais qui servent à la catégorisation du domaine et donnent un accès direct aux documents ;

— résoudre le défaut de couverture de la ressource en s'appuyant sur le vocabulaire de la collection.

Au cours de la propagation, certaines autres fonctionnalités sémantiques peuvent être assurées par notre modèle, *via* l'exploitation de la sémantique implicite ou explicite sur le graphe pondéré :

— la sémantique implicite se manifeste par le phénomène de *co-occurrence* des nœuds du graphe : la prendre en compte permet d'améliorer la précision, le rappel et le classement des résultats, par son impact sur les valeurs d'activation des nœuds, et grâce à la désambiguïsation sémantique du vocabulaire ou de la ressource ;

— la sémantique explicite se manifeste *via* la prise en compte des classes sémantiques (concepts de la couche sémantique), qui permet de traiter la *synonymie*, même en l'absence de ressource dédiée.

4 Expériences

Le but de notre expérimentation est double : tester le passage à l'échelle du modèle en vérifiant qu'on peut le déployer sur une grande collection et étudier les différentes traductions possibles du graphe pondéré ainsi que les différents modes d'interrogation qui en découlent.

La collection *Ohsumed* (Hersh *et al.*, 1994) est une sous collection de MedLine, qui est utilisée pour la tâche de filtrage de TREC-9. Elle rassemble 348 566 références de MedLine, extraites de 270 journaux médicaux sur une période de 5 ans (1987 à 1991) [5]. Nous utilisons dans la suite la partie de la collection datée de 1987, *Ohsumed87* : elle est composée de 54 710

4. C'est à dire indépendante de l'ordre dans lequel les valeurs d'activation des nœuds sont calculées.

5. En général, ces références comportent un titre et un résumé, mais certaines d'entre elles peuvent ne comporter qu'un titre.

documents et la taille moyenne des documents est de 74 56 termes. Nous disposons, toujours dans le cadre de TREC-9, de 63 requêtes avec leurs jugements de pertinence, 3 à 22 documents jugés pertinents par requête.

Des annotations manuelles ont été associées à la collection *Ohsumed* dans le cadre de la tâche de TREC-9, au regard d'une sous-partie du thésaurus MeSH qui a été formalisée en OWL et qui comporte 40 247 concepts (termes MeSH) avec des liens taxonomiques. Nous disposons également d'annotations manuelles des requêtes au regard de MeSH [6].

Nous avons réalisé une annotation automatique du corpus et des requêtes avec l'outil d'annotation METAMAP (Aronson, 2001) qui été conçu pour associer des concepts du méta-thésaurus UMLS à des documents, *a priori* médicaux. Pour expérimenter le passage à l'échelle de notre modèle, nous avons enrichi la plateforme sources ouvertes TERRIER SIR (Bannour & Zargayouna, 2012) avec la plate-forme de gestion et d'analyse des graphes JUNG [7].

Nous avons conduit plusieurs expériences comparatives. Nous avons commencé par distinguer deux modèles de graphe :

— le modèle terme/document (TD), un graphe minimal sans sémantique, avec comme nœuds des termes et des documents, et comme arcs, des liens d'occurrence (termes-documents) pondérés par la fréquence normalisée du terme dans le document [8] ;

— le modèle terme/document/concept (TDC), avec en sus des classes sémantiques et des liens d'annotation (liens documents-concepts avec une valeur de 1) entre les documents et les concepts qui les annotent ; ce modèle a lui-même deux instantiations selon qu'on prend en compte les liens d'annotation manuelle ($TDCman$) ou les liens d'annotation calculés automatiquement ($TDCauto$) [9].

Les résultats sont présentés en termes de R-Précision (R-PREC) qui est la précision après que R documents ont été retrouvés, où R est le nombre de documents pertinents pour la requête considérée. Le nombre d'itérations n'a pas dépassé les 8 itérations par requêtes mais lel temps de traitement des requêtes reste long : nous n'avons pas cherché à ce stade à optimiser l'algorithme de propagation mais à nous assurer de sa robustesse et des fonctionnalités sémantiques qu'il offre.

4.1 Impact de la sémantique

Il s'agit tout d'abord de comparer la RI sans sémantique représentée par le modèle considéré comme la *baseline* et la RI sémantique, en intégrant dans le graphe pondéré une couche sémantique formée des liens terminologiques (termes-documents) et des liens d'annotations (documents-concepts) provenant de l'annotation manuelle ou automatique. Les deux modèles, TD et TDC, sont interrogés par les termes.

Les résultats montrent une différence négligeable (on passe de 18% à 19%, soit un point de pourcentage d'amélioration) mais l'analyse détaillée des requêtes montre des éléments intéressants.

6. L'annotation des requêtes a été réalisée par Gilles Hubert au sein de l'équipe IRIS. (*Information Retrieval & Information Synthesis*), à l'IRIT (Institut de Recherche en Informatique de Toulouse)

7. JUNG (*Java Universal Network/Graph Framework*) accessible à `http://jung.sourceforge.net/`

8. Pour un terme t et un document d : $w(t,d) = w(d,t) = \frac{tf(t,d)}{MaxTf(d)}$

9. Par défaut, ce sont les résultats de l'annotation manuelle qui sont rapportés.

La *co-occurrence des termes et des concepts* apporte une légère amélioration dans l'ordre des documents pertinents retournés pour certaines requêtes. Prenons l'exemple de la requête `REQ9` : « 29 yo female 3 months pregnant. Rh isoimmunization, review topics ». Nous observons une amélioration notable de la R-PREC au niveau de cette requête qui passe de 50% avec le modèle TD à 80% avec le modèle TDC : la co-occurrence des concepts, à la deuxième itération, a permis de renforcer les valeurs d'activation des documents pertinents activés à la première itération. Ces concepts co-occurrents correspondent aux concepts annotant la requête (`#adult, #women #pregnancy`, `#rh_isoimmunization,`) et à d'autres concepts en accord avec la requête comme `#rh-hr_blood-group_system`, `#blood`, `#isoantibody`, etc.

Il y a de *nouveaux documents pertinents découverts au cours de la propagation mais mal classés* en fin de propagation. En fait, on remarque que les documents retrouvés à la première itération restent en haut du classement au cours de la propagation. Ceci montre que les valeurs d'activation s'affaiblissent trop rapidement quand on parcourt le graphe et c'est un point sur lequel le paramétrage du modèle demanderait à être amélioré.

4.1.1 Impact de la densité du graphe

Nous comparons également les modèles $TDCman_{TC}$ et $TDCauto_{TC}$ pour analyser la propagation en fonction de la densité du graphe, les liens étant beaucoup plus nombreux dans le graphe construit à partir de l'annotation automatique. L'interrogation s'est faite par termes et concepts ($_{TC}$).

Le modèle $TDCauto_{TC}$, qui profite de l'annotation automatique de MetaMap, semble plus efficace que le même modèle avec l'annotation manuelle ($TDCman_{TC}$) : on remarque une amélioration de 9% de la R-PREC.

L'examen des résultats requête par requête montre une amélioration pour 27 requêtes (voir le diagramme de la figure 1) qui s'explique principalement par :

— *la résolution des problèmes de couverture* avec la ressource *MeSH* complète utilisée pour l'annotation automatique : pour la requête `REQ58` (« 26 yo woman with mid-thoracic back pain. scheurmann's disease, treatment. ») par exemple, l'annotation manuelle ne permet pas d'annoter le concept `#scheuermann's_disease` que donne l'annotation par MetaMap ;

— *l'exactitude et la précision de l'annotation automatique* : dans la requête `REQ52` (« 60 year old with lung abscess. surgery vs. percutaneous drainage for lung abscess »), on a une amélioration de la *R-PREC* qui passe de 0 à 60% ; cette requête est annotée manuellement par les concepts `#middle_aged`, `#lung`, `#lung_disease`, `#lung_absces` et `#drainage`, c'est-à-dire par des concepts proches de `#lung_absces` (`#lung, #lung_disease`) mais plus généraux ; avec MetaMap, l'annotation de cette requête se limite aux concepts `#Lung Abscess, #Surgery` et `#Drainage`, ce qui minimise le bruit que peuvent introduire des concepts plus généraux.

On note également une dégradation de la R-PREC au niveau de 10 requêtes (voir diagramme de la figure 2). Ces dégradations sont majoritairement dues à des erreurs d'annotation faites par MetaMap pour des problèmes d'ambiguïté. Citons par exemple le cas de la requête `REQ7` (« young wf with lactase deficiency. lactase deficiency therapy options ») où l'annotation du terme ambigu « wf » par MetaMap avec le concept `#Rats, Inbred WF` induit la propagation en erreur.

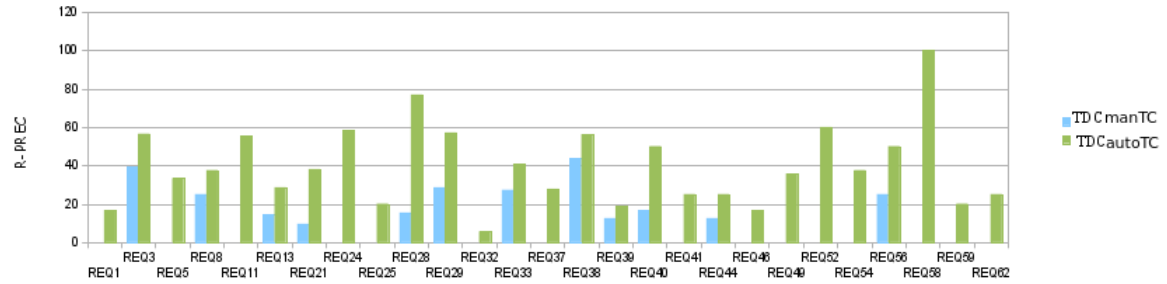


FIGURE 1 – Diagramme de comparaison entre les deux modèles $TDCman_{TC}$ et $TDCauto_{TC}$: amélioration de la R-PREC pour quelques requêtes

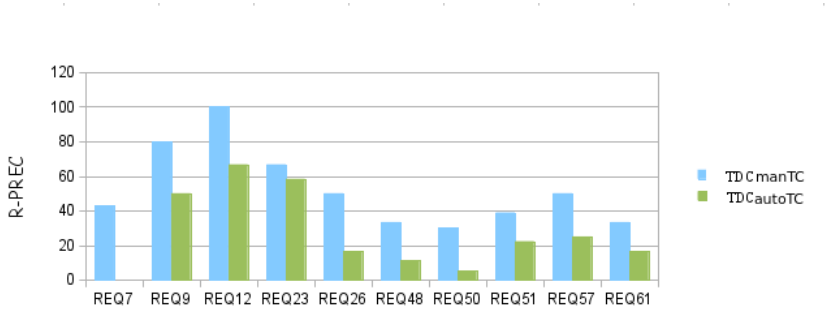


FIGURE 2 – Diagramme de comparaison entre les deux modèles $TDCman_{TC}$ et $TDCauto_{TC}$: dégradation de la R-PREC pour quelques requêtes

L'annotation automatique permet de résoudre certains problèmes de couverture, car l'ontologie MeSH exploitée par MetaMap est complète, ce qui n'est pas le cas pour l'annotation manuelle mais aussi certains problèmes d'ambiguïté et de *term mismatch* entre le vocabulaire de l'utilisateur et la collection. On note également qu'annoter par les concepts les plus spécifiques et uniquement avec ceux-là améliore nettement les performances du système en minimisant le bruit.

5 Conclusion et perspectives

Nous avons présenté dans ce travail un modèle dédié à la recherche d'information sémantique qui donne une vision unifiée des modèles documentaire et sémantique. Différentes configurations du réseau sémantico-documentaire sont possibles, selon les connaissances qu'on choisit de représenter et selon les propriétés symboliques ou numériques qu'on décide d'attribuer aux nœuds et aux arcs du réseau.

Nous avons testé ce modèle sur un corpus médical et montré l'impact des nœuds et des liens pris en compte. Nous avons ainsi pu mesurer l'apport des classes sémantiques ainsi que des liens entre documents et concepts qu'ils soient construits manuellement ou automatiquement.

Ces expériences ont permis de faire une analyse à la fois quantitative et qualitative des résultats : elles montre les fonctionnalités sémantiques qu'apporte le modèle proposé et suggère des pistes d'amélioration. Ainsi, nous avons observé l'impact des problèmes de couverture des ressources. La prise en compte de la co-occurrence termes-concepts permet d'en atténuer l'effet. Nous avons aussi vérifié que nous arrivons à retrouver des documents même si les termes de la requête ne font pas partie du vocabulaire des documents (*term mismatch*).

Certaines erreurs d'annotation introduisent cependant des problèmes d'ambiguïté et dégradent les résultats pour quelques requêtes. Ces erreurs pourraient être corrigées avec l'ajout des liens entre concepts. Cette piste reste à explorer car elle nécessiterait de revoir le mécanisme de contrôle de la propagation qui provoque le relâchement trop rapide des valeurs d'activation avec la distance. Un autre mécanisme de contrôle permettrait de réduire le nombre de nœuds visités et agirait ainsi sur les temps de propagation. D'autres expériences sont à mener pour étudier en détail d'autres paramétrages, concernant par exemple les poids des différents arcs du graphe, mais il faut les conduire de manière systématique pour ne faire varier qu'un paramètre à la fois et analyser son impact sur les résultats globaux. Ces expériences nous permettraient à terme de proposer un calibrage automatique de ces paramètres.

Remerciements

Ce travail a bénéficié partiellement d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'Avenir » portant la référence ANR-10-LABX-0083.

Références

ARONSON A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus : the metamap program. *Proc AMIA Symp*, p. 17–21.

BANNOUR I. & ZARGAYOUNA H. (2012). Une plate-forme open-source de recherche d'information sémantique. In *COnférence en Recherche d'Information et Applications (CORIA)*, p. 167–178.

BANNOUR I., ZARGAYOUNA H. & NAZARENKO A. (2016). Modèle unifié pour la recherche d'information sémantique. In *27es Journées Francophones d'Ingénierie des Connaissances*, Montpellier, France.

BAZIZ M., AUSSENAC-GILLES N. & BOUGHANEM M. (2003). Désambiguisation et Expansion de Requêtes dans un SRI, Etude de l'apport des liens sémantiques . *Revue des Sciences et Technologies de l'Information (RSTI) série ISI*, **8**(4/2003), 113–136.

BHOGAL J., MACFARLANE A. & SMITH P. (2007). A review of ontology based query expansion. *Information Processing and Management*, **43**(4), 866 – 886.

BRIN S. & PAGE L. (1998). The anatomy o f a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web (WWW7)*, p. 107–117, Amsterdam, The Netherlands, The Netherlands : Elsevier Science Publishers B. V.

BROUARD C. (2013). Comparaison du modèle vectoriel et de la pondération tf*idf associée avec une méthode de propagation d'activation. In *CORIA*, p. 1–10, Neuchâtel, France.

CRESTANI F. (1997). Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, **11**(6), 453–482.

CRESTANI F. (2000). Exploiting the similarity of non-matching terms at retrieval time. *Information Retrieval*, **2**(1), 27–47.

DINH D. & TAMINE L. (2010). Vers un modèle d'indexation sémantique adapté aux dossiers médicaux de patients. *Conférence francophone en Recherche d'Information et Applications, CORIA 2010*, p. 325–336.

FERNÁNDEZ M., CANTADOR I., LÓPEZ V., VALLET D., CASTELLS P. & MOTTA E. (2011). Semantically enhanced information retrieval : an ontology-based approach. *Web Semantics : Science, Services and Agents on the World Wide Web*, **9**(4), 434–452.

GUARINO N., MASOLO C. & VETERE G. (1999). Ontoseek : Using large linguistic ontologies for accessing on-line yellow pages and product catalogs. *National Research Council, LADSEBCNR*.

HAMADAN H., ALBITAR S., BELLOT P., ESPINASSE B. & FOURNIER S. (2012). Lsis at trec 2012 medical track – experiments with conceptualization, a dfr model and a semantic measure. In *The Twenty-First Text REtrieval Conference (TREC 2012) Notebook*, volume Special Publication, p. 12 p., Gaithersburg (USA).

HERSH W., BUCKLEY C., LEONE T. & HICKAM D. (1994). Ohsumed : An interactive retrieval evaluation and new large test collection for research. In *SIGIR'94*, p. 192–201 : Springer.

JIANG X. & TAN A.-H. (2006). Ontosearch : A full-text search engine for the semantic web. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, p. 1325–1330 : AAAI Press.

KHAN L. R. (2000). *Ontology-based Information Selection*. PhD thesis, Faculty of the Graduate School, University of Southern California.

MIMOUNI N., NAZARENKO A., PAUL È. & SALOTTI S. (2014). Towards graph-based and semantic search in legal information access systems. In *Legal Knowledge and Information Systems - JURIX*, volume 271, p. 163–168.

NARULA G. S. & JAIN V. (2014). Information retrieval (IR) through semantic web (SW) : an overview. *CoRR*, **abs/1403.7162**.

QUILLIAN M. R. (1968). Semantic memory. In *Semantic information processing*. Cambridge : MIT Press.

ROCHA C., SCHWABE D. & ARAGAO M. P. (2004). A hybrid approach for searching in the semantic web. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, p. 374–383, New York, NY, USA : ACM.

SALTON G. (1968). *Automatic Information Organization and Retrieval*. McGraw Hill Text.

SCHUMACHER K., SINTEK M. & SAUERMANN L. (2008). Combining fact and document retrieval with spreading activation for semantic desktop search. In *The Semantic Web : Research and Applications*, volume 5021 of *Lecture Notes in Computer Science*, p. 569–583. Springer Berlin Heidelberg.

STETINA J., KUROHASHI S. & NAGAO M. (1998). General word sense disambiguation method based on a full sentential context. In *Proceedings of COLING-ACL workshop, Usage OF Wordnet in natural langage processing*.

WANG H., TRAN T., LIU C. & FU L. (2011). Lightweight integration of ir and db for scalable hybrid search with integrated ranking support. *Web Semant.*, **9**(4), 490–503.

ZARGAYOUNA H., ROUSSEY C. & CHEVALLET J.-P. (2015). Recherche d'information sémantique : état des lieux. *Traitement Automatique des Langues*, **56**(3), 49–73.

ZHANG L., LIU Q., ZHANG J., WANG H., PAN Y. & YU Y. (2007). Semplore : An ir approach to scalable hybrid query of semantic web data. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007), Busan, South Korea*, volume 4825 of *LNCS*, p. 645–658, Berlin, Heidelberg : Springer Verlag.

Contribution à la recherche de vérité : modèles exploitant des règles d'association extraites d'une base de connaissances

Valentina Beretta[1], Sylvie Ranwez[1], Sébastien Harispe[1] et Isabelle Mougenot[2]

[1] LGI2P de l'école des mines d'Alès, Site de Nîmes, Parc G. Besse, F-30 035 Nîmes
{prenom.nom}@mines-ales.fr
[2] UMR Espace-Dev, Université de Montpellier, Rue JF. Breton, Montpellier
isabelle.mougenot@umontpellier.fr

Résumé : Pour contrer les dangers de la désinformation, un domaine de recherche a émergé ces dernières années : la détection de vérité sur le Web. Héritière de la vérification de faits (*fact checking*) d'une part et des techniques de fusion de données d'autre part, la détection de vérité analyse les *déclarations* (i.e. <*sujet, prédicat, valeur*>) diffusées par plusieurs *sources* sur un sujet donné, et tente de déterminer parmi toutes ces déclarations, celle qui constitue un *fait* (une *vérité*). Nous avons récemment montré que la prise en compte d'axiomes fournis par une ontologie de domaine, et en particulier certaines relations transitives, constitue une réelle plus-value et permet d'améliorer la performance des approches existantes. Dans cet article, nous enrichissons nos travaux précédents en étendant l'analyse à d'autres types de relations, pouvant participer à l'identification de règles pour renforcer la confiance dans certaines déclarations et ainsi améliorer l'identification de vérité. A partir d'une base de connaissances, un ensemble de règles est extrait. Un coefficient *propulseur* (*booster*) est alors calculé pour renforcer certaines déclarations. Les premiers résultats de l'évaluation montrent une plus-value par rapport aux approches traditionnelles.

Mots-clés : Détection de vérité, Ontologies, Web sémantique, Détection de règles, Raisonnement.

1 Introduction

Après une certaine période d'euphorie engendrée par l'accès et la diffusion à très grande échelle d'un grand nombre d'informations aussi diverses que peuvent l'être nos différentes activités humaines : relation sociales, activités professionnelles, engagements associatifs et/ou politiques, création artistiques ou photographiques, … l'heure est à la prudence. Les mises en garde sont de plus en plus soutenues auprès des personnes les plus "vulnérables" et en particulier des jeunes générations, afin d'éviter la propagation d'informations fausses et l'adhésion à certaines idéologies qui constitueraient une menace pour nos sociétés. Le site Politifact[1] analyse ainsi depuis plusieurs années les discours des responsables politiques américains afin de déterminer leur part de vérité et de mensonge. En France, le journal Le Monde propose un outil de vérification de la fiabilité des sources (Décodex[2]). Dans les deux cas, ce sont des acteurs humains (journalistes principalement) qui analysent les contenus et composent des synthèses qui sont restituées au grand public. Mais le volume d'informations est trop important pour être traité de façon exhaustive et des approches automatisées sont nécessaires pour les assister dans leur tâche. Pour contrer les dangers de la désinformation, un nouveau domaine de recherche a émergé ces dernières années : la détection de vérité sur le Web (*Truth finding*). Héritière de la vérification de faits (*fact checking*) d'une part et des techniques de fusion de données d'autre part, la détection de vérité analyse les *déclarations*

[1] www.politifact.com
[2] http://www.lemonde.fr/verification/

diffusées par plusieurs sources sur un sujet donné, et tente de déterminer parmi toutes ces déclarations, celle qui constitue un *fait* (une *vérité* objective[3]). Cette étape est particulièrement importante lorsque l'on souhaite enrichir des bases de connaissances à partir de processus d'extraction automatique complexes faisant intervenir plusieurs extracteurs (sources), afin de constituer un support, par exemple, pour l'aide à la décision.

Les techniques actuelles de recherche de vérité se basent principalement sur un postulat : les sources qui ont diffusé majoritairement des déclarations vraies sont estimées comme étant *fiables* et avec une forte propension à dire la *vérité*. La *confiance* dans les informations qu'elles diffusent est alors considérée comme d'autant plus élevée (Li et al., 2015). Un processus itératif est utilisé afin de calculer ces degrés de fiabilité et de confiance et ainsi déterminer les déclarations qui traduisent des *faits* (vérités). Nos travaux s'inscrivent dans cette veine et intègrent les ontologies de domaine au processus de détection de vérité afin de renforcer la confiance associée à certaines affirmations. Ainsi, dans (Beretta et al., 2016), nous avons proposé de prendre en compte certaines relations transitives d'une ontologie qui définissent un ordre partiel sur les valeurs pour en confirmer certaines et tenter d'identifier les valeurs *vraies*. Nous considérions alors seulement une portion réduite de l'ontologie, essentiellement au travers de l'exploitation partielle des définitions de classes contenues dans la T-Box. Plus précisément, notre approche se concentrait sur les ordres partiels des ressources formés par la structuration des classes (*e.g.* **`subClassOf`**), le typage des ressources (*e.g.* **`type`**) et d'éventuels liens entre ressources fournis par des prédicats transitifs supplémentaires (*e.g.* **`partOf`**). Dans cette contribution, nous souhaitons aller plus loin en intégrant une analyse plus large de la A-Box, afin de prendre en compte tous les types de relation et les *faits* qui la composent. En effet, en étudiant les cooccurrences entre ces faits, il est possible d'identifier des motifs qui peuvent être ensuite utilisés pour conforter notre jugement *a priori* sur certaines déclarations. Prenons un exemple. Le fait qu'une personne soit née en Espagne est fréquemment associé au fait que cette même personne parle espagnol. Ainsi, si l'on recherche le lieu de naissance de Pablo Picasso sachant qu'il parle espagnol, lors de la recherche de vérité le système pourrait renforcer la confiance dans les déclarations qui proposent une valeur correspondant à l'Espagne ou à des valeurs plus génériques. C'est l'idée que nous explorons dans cette étude.

La section suivante présente le contexte de notre étude, certaines notations et les travaux existants. La section 3 formalise notre approche et détaille l'intégration de règles dans la procédure itérative qui calcule alternativement la fiabilité des sources et la confiance dans les faits pour déterminer les valeurs vraies. La section 4 discute les premiers résultats obtenus et enfin la section 5 conclut cet article et ouvre de nombreuses perspectives de recherche.

2 Positionnement et état de l'art : bases de connaissances et règles d'association

Notre étude s'appuie sur des graphes de connaissances tels que DBpedia (Auer et al., 2007) où les nœuds correspondent à des entités de différents types et les arcs correspondent à des relations, également de différents types, entre ces entités. Nous proposons d'améliorer le processus de recherche de vérité en exploitant les informations contenues dans un tel graphe.

2.1 Bases de connaissances et recherche de vérité

De façon plus formelle, appelons KB une base de connaissances supposée n'être composée que de *faits* (*vérités* objectives) représentés par un ensemble de triplets RDF de la forme <*sujet, prédicat, objet*>. Plusieurs types de prédicats et d'entités (*sujets* et *objets*) sont autorisés. L'objectif ultime de notre approche concerne l'enrichissement de cette base de

[3] Suivant la distinction qui est faite en philosophie entre *vérité de fait* et *vérité de raison*, la définition de la *vérité* considérée ici est la *vérité de fait* : Enoncé qui correspond au réel qu'il décrit (*vérité contingente* selon Leibniz ou *relation de faits* selon Hume).

connaissances afin de constituer un meilleur support pour l'aide à la décision. Chaque élément rajouté à la base doit donc être fiable.

Une paire (*sujet, prédicat)* est appelée une *description*[4] et représente une propriété particulière d'une entité sujet. La valeur associée à cette propriété est représentée par le singleton {*valeur*}. Il est à noter que la recherche de vérité envisagée ici ne concerne que des prédicats fonctionnels, c'est à dire pour lesquels une seule valeur est possible (e.g. une personne ne peut être née qu'à un seul endroit). Lors d'un processus d'extraction de connaissances (par exemple à partir d'analyse de textes), plusieurs sources d'information[5] peuvent proposer des valeurs différentes et contradictoires pour une même description. Ces propositions, également représentées par des triplets <*sujet, prédicat, valeur*>, sont appelées *déclarations* tant qu'elles ne sont pas validées, i.e. tant que l'on n'a pas identifié la valeur *vraie*. Il est, en effet, nécessaire de déterminer quelle est cette valeur *vraie*, parmi celles proposées, afin de constituer un nouveau *fait* qui pourra être intégré à la base.

Nous ne détaillons pas ici l'état de l'art concernant la détection de vérité. Le lecteur intéressé pourra se reporter à (Berti-Équille & Borge-Holthoefer, 2015) pour un état de l'art approfondi. Dans les approches traditionnelles de détection de vérité, un processus itératif calcule alternativement la *confiance* dans les déclarations et la *fiabilité* des sources afin de déterminer quelle est la valeur vraie la plus probable. La détection de vérité peut également exploiter différentes dépendances : entre les sources (Blanco et al., 2010; Dong et al., 2010; Pochampally et al., 2014; Qi et al., 2013), entre les valeurs (Yin et al., 2008) ou entre les descriptions (Meng et al., 2015; D. Wang et al., 2015; S. Wang et al., 2015). Ces modèles ne fournissent bien souvent qu'un score numérique.

Bien que n'ayant jamais été utilisées dans le domaine de la détection de vérité, il est également possible d'exploiter les cooccurrences par l'identification de règles d'association (Maimon & Rokach, 2005). Cette solution exprime une sémantique et permet une interprétation, c'est pourquoi nous avons choisi d'explorer cette voie. Dans ce qui suit, nous souhaitons utiliser un coefficient propulseur (*booster*), calculé à partir de l'identification de cooccurrences récurrentes entre différents faits afin de renforcer la confiance dans certaines valeurs pendant le processus de détection de vérité.

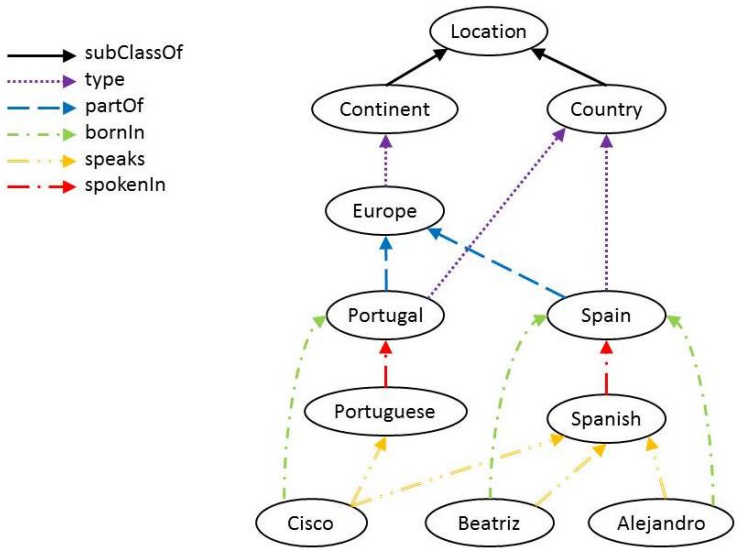


FIGURE 1 – Extrait d'une base de connaissances.

Prenons l'exemple très simplifié représenté dans la **FIGURE 1**. Une analyse de la base permet de déduire que la majorité des personnes qui parlent espagnol sont nées en Espagne.

[4] Nous employons le terme *description* comme traduction de *data item* couramment utilisé dans la littérature anglaise.

[5] Ici "source d'information" est employé au sens large : il peut d'agir d'un site Internet, d'une base de données, d'une personne (via l'analyse de ses écrits…). On simplifiera le propos par la suite en ne parlant que de "source".

Cette observation peut être prise en compte dans un processus de recherche de vérité concernant le lieu de naissance de Pablo Picasso, par exemple. Si on observe que Pablo Picasso parle couramment espagnol, la confiance attribuée à la déclaration <Picasso, bornIn, Spain> doit être renforcée, ainsi que les déclarations qui contiennent des valeurs plus génériques[6]. Ce renforcement est apporté par le coefficient propulseur qui représente le degré de soutien (la caution) apporté pour cette déclaration par les informations contenues dans KB. Ainsi le postulat de base du processus de recherche de vérité présenté en introduction sera modifié et nous considèrerons désormais que les *faits* (*vérités*) sont des *déclarations* proposées par des sources fiables et/ou qui sont renforcées par un coefficient *booster* élevé en considération de règles d'association extraites de KB. Comme dans les approches traditionnelles, la fiabilité d'une source dépendra, quant à elle, du nombre de vérités qu'elle a proposées. Il est à noter que les motifs récurrents n'ont pas tous le même degré d'expressivité et ne doivent donc pas avoir le même impact sur le processus de détection de vérité. L'influence du coefficient *booster* sur le calcul de confiance dans une déclaration sera donc paramétrable afin d'accorder plus d'importance à la fiabilité des sources ou au contraire à l'information contenue dans KB en fonction du contexte et/ou de la qualité de la base.

2.2 Détection de règles d'association : principes et mise en œuvre

Comme mentionné dans la synthèse sur les règles d'association présentée dans (Maimon & Rokach, 2005), il est difficile d'avoir une vue exhaustive des travaux dans ce domaine. Pour des applications en lien avec le Web sémantique, on peut se référer à (Galárraga et al., 2015; Z. Wang & Li, 2015). Certains problèmes sont particuliers à ce contexte : la quantité de données, l'assomption du monde ouvert et les données manquantes (Quboa & Saraee, 2013).

Dans la suite, la notation utilisée pour les règles sera celle de Datalog. Une règle est une implication d'un ensemble d'atomes reliés par un opérateur de conjonction, appelé corps (aussi appelé antécédent ou prémisse), vers un autre ensemble appelé tête (conséquence). Formellement, la règle R pourra s'écrire :

$R: B_1 \wedge B_2 \wedge \ldots \wedge B_n \Rightarrow H$ qui est équivalent à $R: \vec{B} \Rightarrow H$.

Dans notre approche, nous considérons uniquement des clauses de Horn, c'est-à-dire qui n'ont qu'un singleton dans la *tête*. Ici, un atome est assimilé à une *déclaration* constituée d'un prédicat défini et de sujet et objet qui peuvent être des variables. Pour simplifier l'écriture, une déclaration constituée d'un triplet <*s*, *p*, *v*> sera notée : $p(s, v)$. L'identification des règles par l'analyse de la base de connaissances est réalisée avec AMIE+ (Galárraga et al., 2015). Ces règles ont notamment deux propriétés : i) elles sont connectées (chaque atome est transitivement connecté avec les autres atomes), ii) elles sont fermées, i.e. elles contiennent des variables fermées, c'est-à-dire qui apparaissent au minimum deux fois dans la règle.

Plusieurs métriques ont été proposées pour évaluer la qualité d'une règle, dont les plus répandues sont le *support* et la *confiance* (Feno, 2007).

Le *support* indique la proportion d'entités vérifiant à la fois le corps et la tête de la règle. Dans notre contexte de raisonnement en monde ouvert, nous utilisons la définition de (Galárraga et al., 2015). Pour la règle $R: \vec{B} \Rightarrow H$ où H est composé d'une seule déclaration $p(s, v)$, le support est calculé de la façon suivante :

$$supp(R) := supp\left(\vec{B} \Rightarrow p(s, v)\right) := \#(s, v): \exists x_1, \ldots, x_i: \vec{B} \wedge p(s, v) \quad (1)$$

[6] Ce dernier point ne sera pas discuté ici puisqu'il rejoint la contribution de (Beretta et al., 2016).

où $x_1, \dots, x_i$ représentent les variables contenues dans les atomes de $\vec{B}$ exceptées s et v. Si l'on considère l'exemple présenté dans la TABLE 1, le *support* de la règle $r: livesIn(s,v) \Rightarrow bornIn(s,v)$ serait égal à 1, étant donné qu'on ne trouve dans la base de connaissances que la paire de déclarations $livesIn(Adam, Paris)$ et $bornIn(Adam, Paris)$ qui la vérifie.

TABLE 1 – *Extraits de faits de la base de connaissances. Les prédicats sont notés en tête de colonne et chaque cellule contient un couple (entité, valeur) pour ce prédicat.*

livesIn	*bornIn*
(Adam, Paris)	(Adam, Paris)
(Adam, Caen)	(Luca, Rome)
(Luca, Milan)	(Carl, London)
(Bob, Lugano)	

La *confiance*, quant à elle, indique la proportion d'entités vérifiant la tête, parmi celles qui vérifient le corps. Cette mesure, comprise entre 0 et 1 n'est pas sensible à la taille des données. On peut la calculer de la façon suivante :

$$conf\left(\vec{B} \Rightarrow p(s,v)\right) := \frac{supp\left(\vec{B} \Rightarrow p(s,v)\right)}{supp(\vec{B})} := \frac{\#(s,v): \exists x_1,\dots,x_i: \vec{B} \wedge p(s,v)}{\#(s,v): \exists x_1,\dots,x_i: \vec{B}} \quad (2)$$

Dans notre exemple, $conf\left(livesIn(s,v) \Rightarrow bornIn(s,v)\right) = \frac{1}{4}$. Cette mesure de confiance a été définie dans un contexte de raisonnement en monde fermé, où l'on considère comme fausses les déclarations qui ne sont pas exprimées dans la base. Or dans le contexte du Web sémantique, celui qui nous concerne dans cette étude, c'est l'hypothèse d'un monde ouvert qui est envisagée selon les principes qui ont cours dans les Logiques de Description. C'est pour cela que les auteurs de (Galárraga et al., 2015) ont introduit la mesure de *PCA confidence* qui repose sur l'hypothèse de complétude partielle (PCA pour *Partial Completeness Assumption*) qui considère que si la base de connaissances contient au moins un triplet qui concerne une description, i.e. une paire (*sujet*, *prédicat*), alors toutes les valeurs possibles pour cette description sont connues. Autrement dit, si une description n'apparait jamais dans la base, elle n'est considérée ni comme étant vraie, ni comme étant fausse. La mesure de *PCA confidence* se calcule comme suit :

$$conf_{PCA}\left(\vec{B} \Rightarrow p(s,v)\right) := \frac{supp\left(\vec{B} \Rightarrow p(s,v)\right)}{\#(s,v): \exists x_1,\dots,x_i,y: \vec{B} \wedge p(s,y)} \quad (3)$$

Dans notre exemple, $conf_{PCA}\left(livesIn(s,v) \Rightarrow bornIn(s,v)\right) = \frac{1}{3}$, puisqu'on rencontre une fois la règle ($livesIn(Adam, Paris)$ et $bornIn(Adam, Paris)$) et trois fois une variable impliquée dans la prémisse de cette règle ($livesIn(Adam, Paris)$, $livesIn(Adam, Caen)$ et $livesIn(Luca, Milan)$) .

Le support et la confiance représentent deux caractéristiques d'une même règle. Dans notre cas il est important de considérer ces deux mesures. En effet, dans certains cas une règle R pourra avoir une mesure de $conf_{PCA}(R) = 1$ et une mesure de support $supp(R) = 2$ alors qu'une autre règle R' pourra avoir également une mesure $conf_{PCA}(R') = 1$ mais un support $supp(R') = 100$. Dans ce cas-là, on préfèrera se baser sur la règle R' qui a été observée un plus grand nombre de fois que la règle R. Dans le même ordre d'idée, choisir une règle uniquement parce qu'elle a une mesure de confiance bien supérieure aux autres règles n'a de sens que si elle a été observée un grand nombre de fois. Nous avons donc choisi une fonction d'agrégation qui permet de considérer simultanément ces mesures dans un même indicateur. Nous nous sommes notamment basés sur le modèle d'agrégation proposé dans (Jean, Harispe,

Ranwez, Bellot, & Montmain, 2016). Ce modèle a été adapté à notre contexte et résulte dans la formulation suivante. Soit une règle R, son support $supp(R)$ et sa confiance $conf_{PCA}(R)$, le score obtenu par agrégation de ces différentes caractéristiques est donné par :

$$score(R) = \left(1 - \frac{1}{supp(R)}\right) conf_{PCA}(R) \quad (2)$$

Ainsi en pondérant la mesure de confiance dans une règle par les occurrences de cette règle, on accorde plus de confiance dans les règles qui sont les plus fréquentes (avec un support élevé).

3 Utilisation de règles pour la détection de vérité

Ici, les règles d'association sont identifiées grâce à AMIE+ (Galárraga et al., 2015) ainsi que les mesures de *support* et de *PCA confidence*. Disposant de ces informations, l'approche proposée consiste à adapter les méthodes existantes en intégrant un coefficient propulseur ($booster$), que nous appelons $boost(d)$, dans la procédure itérative de détection de vérité. Ce facteur a une influence directe sur le calcul de confiance dans une déclaration d et une influence indirecte sur le calcul de fiabilité des sources.

TABLE 2 – *Synthèse des notations utilisées dans notre approche.*

symbole	*signification*
$d \in D$	Une déclaration appartenant à l'ensemble de toutes les déclarations
$s \in S$	Une source appartenant à l'ensemble des sources
D^s	L'ensemble des déclarations faites par la source s
S^d	L'ensemble des sources qui proclament une déclaration d
$t^i(s)$	Calcul de la fiabilité (*trustworthiness*) d'une source à l'étape i
$c^i(d)$	Calcul de la confiance dans une déclaration à l'étape i

En utilisant les notations synthétisées dans la TABLE 2, une adaptation de la méthode *Sums* (Pasternack & Roth, 2010), peut être modélisée comme suit pour tenir compte de l'information amenée par les règles :

$$t^i(s) = \frac{1}{max_{s' \in S}\left(\sum_{d' \in D^{s'}} C^{i-1}(d')\right)} \sum_{d \in D^s} c^{i-1}(d) \quad (5)$$

$$c^i(d) = \frac{1}{norm_d}\left((1-\gamma) confidence_{basic}(d) + \gamma . boost(d)\right) \quad (6)$$

avec $\gamma \in [0,1]$ un poids qui représente l'influence relative accordée aux sources et à la base de connaissances ; $confidence_{basic}$ une fonction de D dans [0,1] qui représente la confiance donnée par les sources à une déclaration ; et $boost$ une fonction de D dans [0,1] qui représente la confiance dans une déclaration provenant de l'application des règles identifiées grâce à l'analyse de la base de connaissances.

Le paramètre γ dépend du contexte et sera fixé en fonction de la stratégie choisie. Dans l'évaluation qui sera présentée dans la section 4, plusieurs valeurs seront considérées et leur impact sera discuté.

Le calcul de $confidence_{basic}(d)$ est réalisé comme dans la méthode *Sums* :

$$confidence_{basic}(d) = \frac{\sum_{s \in S^d} t^i(s)}{max_{d' \in D} \sum_{s' \in S^{d'}} t^i(s')} \quad (7)$$

où l'on retrouve au numérateur la somme des fiabilités associées à toutes les sources qui émettent une déclaration et un facteur de normalisation au dénominateur, i.e. la confiance maximale associée à une déclaration.

Le facteur propulseur, *booster*, cherche à synthétiser les informations données par toutes les règles obtenues pour chaque déclaration. Par exemple pour $bornIn$, à partir du graphe de la FIGURE 1, on peut obtenir les règles suivantes :

- $speaks(x,z) \wedge officialLanguage(y,z) \Rightarrow bornIn(x,y)$
- $speaks(x,Spanish) \wedge officialLanguage(Spain,Spanish) \Rightarrow bornIn(x,Spain)$.

Comme nous l'avons montré dans la section 2.2, chaque règle peut être évaluée par un score unique. Les scores des différentes règles qui concernent une même déclaration peuvent ainsi être agrégés. En effet, notre objectif reste bien de renforcer la confiance dans certaines déclarations en utilisant tous les motifs identifiés.

Soit une déclaration $d = p(s,o)$, une base de connaissances KB et un ensemble de règles $R = \{r: B_1 \wedge \ldots \wedge B_n \Rightarrow p'(x,y)\}$ extraites à partir de KB, nous considérons que le facteur propulseur doit être fonction du pourcentage de règles qui sont vérifiées par la déclaration considérée. Pour chaque déclaration, l'ensemble des règles R_d à considérer (règles éligibles) est un sous-ensemble des règles extraites : $R_d \subset R$. Ces règles doivent répondre à certaines contraintes : contenir le prédicat p dans la tête de la règle et avoir un corps composé uniquement d'atomes valides (i.e. contenus dans KB). Formellement $R_d = \{r: B_1 \wedge \ldots \wedge B_n \Rightarrow p'(x,y) \in R | (p' = p) \wedge T(B_1 \wedge \ldots \wedge B_n) = 1\}$ avec $T(B) = 1$ (resp. $= 0$) une fonction qui indique que le corps de la règle est vérifié (resp. n'est pas vérifié). Le facteur propulseur peut alors être défini comme suit.

$$\text{boost(d)} = \left(1 - \frac{1}{1+\sum_{\text{r}\in \text{R}_\text{d}} \text{score(r)}}\right) \frac{\sum_{\text{r}\in \text{R}_\text{d}^\text{T}} \text{score(r)}}{\sum_{\text{r}\in \text{R}_\text{d}} \text{score(r)}} \quad (8)$$

où $\text{R}_\text{d}^\text{T} = \{r: B_1 \wedge \ldots \wedge B_{|r|} \Rightarrow p'(x,y) \in R_d : T(r) = 1\}$ et où $T(r) = 1$ (resp. $= 0$) représente le fait que la règle r soit vérifiée (resp. fausse). Autrement dit, l'ensemble des règles éligibles est composé des règles qui sont vérifiées au moins par une instanciation de leur corps, i.e. contenant le même sujet pour le prédicat considéré.

Ce facteur propulseur est compris entre 0 et 1.

Prenons l'exemple de la *TABLE 3*. Si l'on considère les deux règles suivantes :

- $R^1: speaks(x,z) \wedge officialLanguage\ (y,z) \Rightarrow bornIn(x,y)$
- $R^2: resident(x,France) \Rightarrow bornIn(x,France)$

où le score de R^1 est de 0,55 et celui de R^2 $est\ de$ 0,75. Le coefficient propulseur pour d_1 est de 0.245 car les deux règles sont éligibles, mais seule R^1 est vérifiée ce qui donne : (1-1/(1+0.55+0.75))*(0.55/(0.55+0.75)). Par contre, pour d_2 le score sera de 0 car même si les deux règles sont éligibles, aucune n'est vérifiée.

TABLE 3 – *Eléments d'une base de connaissances.*

Sources/origine	*Déclarations*
s_1	$d_1 = bornIn(Picasso, Spain)$
s_2	$d_2 = bornIn(Picasso, UK)$
KB	$d_3 = officialLanguage(Spain, Spanish)$
KB	$d_4 = speaks(Picasso, Spanish)$
KB	$d_5 = resident(Picasso, France)$

Nous avons implémenté quatre modèles différents à partir de la méthode *Sums*. La méthode *Sums* dite traditionnelle (M_1) est celle qui est proposée par les auteurs de (Pasternack

& Roth, 2010). La méthode M_2 consiste à intégrer à *Sums* la prise en compte les règles identifiées (après analyse de la A-Box) lors du calcul de la confiance dans une déclaration (comme décrit ci-dessus). La méthode M_3 est la méthode présentée dans (Beretta et al., 2016) qui consiste à tenir compte uniquement des relations transitives en plus de la méthode *Sums* et enfin la méthode M_4 consiste à tenir compte à la fois des relations définies dans l'ontologie et des règles identifiées par l'analyse de la A-Box dans le calcul de confiance associée aux déclarations. La *TABLE 4* synthétise ces différents modèles et les équations associées respectivement au calcul de la confiance dans les déclarations et au calcul de la fiabilité des sources dans le processus itératif de recherche de vérité.

TABLE 4 – *Récapitulatif des différents modèles utilisés pour la recherche de vérité.*

M_1 – *Sums* traditionnel
$t^i(s) = \frac{1}{max_{s' \in S}\left(\sum_{d' \in D^{s'}} c^{i-1}(d')\right)} \sum_{d \in D^s} c^{i-1}(d)$
$c^i(d) = confidence_{basic}(d) = \frac{\sum_{s \in S^d} t^i(s)}{max_{d' \in D}\left(\sum_{s' \in S^{d'}} t^i(s')\right)}$
M_2 – *Sums* traditionnel + prise en compte des règles d'association
$t^i(s) = \frac{1}{max_{s' \in S}\left(\sum_{d' \in D^{s'}} c^{i-1}(d')\right)} \sum_{d \in D^s} c^{i-1}(d)$
$c^i(d) = \frac{1}{norm_d}[(1-\gamma) confidence_{basic}(d) + \gamma . boost(d)]$
M_3 – *Sums* traditionnel + propagation en fonction des relations transitives de l'ontologie
$t^i(s) = \frac{1}{max_{s' \in S}\left(\sum_{d' \in D^{s'}} c^{i-1}(d')\right)} \sum_{d \in D^s} c^{i-1}(d)$
$c^i(d) = adaptedConfidence(d) = \frac{\sum_{s \in S^{d+}} t^i(s)}{max_{d' \in D} \sum_{s' \in S^{d'}} t^i(s')}$ $avec\ S^{d+} = S^d \cup \{s \in S^{d'} : d' \in D \wedge d' \preccurlyeq d\}$
M_4 – *Sums* traditionnel + propagation en fonction des relations transitives + Règles
$t^i(s) = \frac{1}{max_{s' \in S}\left(\sum_{d' \in D^{s'}} c^{i-1}(d')\right)} \sum_{d \in D^s} c^{i-1}(d)$
$c^i(d) = \frac{1}{norm_d}[(1-\gamma) adaptedConfidence(d) + \gamma . boostProp(d)]$ avec $boostProp(d) = \left(1 - \frac{1}{\sum_{r \in R_{d+}} score(r)}\right) \frac{\sum_{r \in R^T_{d+}} score(r)}{\sum_{r \in R_{d+}} score(r)}$ et $R_{d+} = R_d \cup \{R_{d'} \in R : d' \in D \wedge d' \preccurlyeq d\}$

Il est à noter que $boostProp(d)$ est un coefficient calculé à partir du coefficient *boost* qui provient de l'application des règles d'association mais auquel on applique une propagation. En effet, soit une règle dont la tête est égale à $H = p(s, o)$, une telle règle se vérifie et donc confirme toutes les règles plus génériques au regard de l'ontologie de domaine (i.e. les règles qui impliquent des concepts plus génériques que ceux de la règle considéré). Autrement dit, le facteur de *boost* correspondant doit être propagé à tous les *ancêtres*. Ce facteur permet

d'assurer la monotonie de la fonction de confiance associée aux déclarations. En effet, la confiance $c(d)$ dans une déclaration d telle que $d' \preccurlyeq d$ doit être supérieure ou égale à la confiance $c(d')$ associée à la déclaration d'.

4 Discussion des résultats

Les expérimentations qui suivent ont été réalisées sur le corpus exploité dans (Beretta et al., 2016), disponible à l'adresse https://doi.org/10.6084/m9.figshare.4616071. Ce jeu de test a été réalisé à partir de l'extraction du prédicat **`dbpedia-owl:birthPlace`** dans DBpedia (version 2015-04) qui permet de disposer du lieu de naissance des personnes connues. Trois jeux de test ont été générés (EXP, LOW_E et UNI) qui diffèrent par la stratégie de sélection des valeurs vraies (plus ou moins spécifiques) – cf. (Beretta et al., 2016) pour plus de détail.

Dans nos expérimentations, nous avons sélectionné uniquement les règles détectées par AMIE+ qui ont une couverture supérieure à 0.012 pour la tête. Nous restreignons ainsi le nombre de règles considéré à 62. Dans chaque cas, l'identification de vérité est réalisée par un processus itératif où les calculs de confiance dans les déclarations et de fiabilité des sources sont ceux présentés dans la ***TABLE 4***. La sélection de la valeur vraie est ensuite réalisée par un algorithme glouton – cf. (Beretta et al., 2016).

TABLE 5 – *Synthèse des résultats obtenus avec les modèles de détection de vérité M_2 et M_4 appliqués aux trois types de corpus. Différentes valeurs de γ ont été testées. Les valeurs indiquées en rouge indiquent les plus mauvais résultats et les résultats en gras les meilleurs.*

		Jeu de données EXP			***Jeu de données LOW_E***			***Jeu de données UNI***		
	γ	n_{vrai}	n_{gen}	n_{faux}	n_{vrai}	n_{gen}	n_{faux}	n_{vrai}	n_{gen}	n_{faux}
M_2 *Sums* + Règles	0*	*0,6267*	*0,1136*	*0,2596*	*0,1726*	*0,1896*	*0,6378*	*0,0335*	*0,1953*	*0,7712*
	0,25	**0,6278**	0,1433	0,2289	0,2085	0,2168	0,5746	0,0737	0,2316	0,6947
	0,50	0,6036	0,1919	0,2045	0,2200	0,2663	0,5137	0,0983	0,2814	0,6203
	0,75	0,5811	0,2404	0,1785	0,2310	0,3125	0,4565	0,1167	0,3284	0,5549
	0,90	0,5944	0,2567	**0,1489**	**0,2507**	0,3439	0,4054	0,1274	0,3695	0,5031
	1,00	0,1578	**0,3205**	0,5217	0,1279	**0,4751**	**0,3970**	**0,1293**	**0,4388**	**0,4318**
M_4 *Sums* adapté + Règles	0*	*0,8955*	***0,0033***	*0,1012*	*0,6645*	*0,0048*	*0,3307*	*0,4938*	*0,0055*	*0,5007*
	0,25	0,8995	0,0033	0,0971	0,6927	0,0049	0,3024	0,5371	0,0057	0,4572
	0,50	0,8983	0,0033	0,0983	0,6969	0,0049	0,2981	0,5454	0,0057	0,4489
	0,75	0,8997	0,0033	0,0970	0,7040	0,0050	0,2911	0,5547	0,0057	0,4396
	0,90	**0,9041**	0,0034	**0,0925**	**0,7142**	0,0050	**0,2808**	**0,5666**	0,0058	**0,4276**
	1,00	0,1018	**0,1545**	0,7437	0,1006	**0,1540**	0,7454	0,1002	**0,1546**	0,7453

Nous avons évalué la précision de notre approche. L'ensemble des résultats est présenté dans la ***TABLE 5***. Nous avons analysé la proportion de valeurs vraies retournées par les différentes méthodes et qui correspondent à des valeurs attendues (pour chaque description (*sujet*, *prédicat*), la valeur attendue est contenue dans un corpus de référence) – n_{vrai}. La proportion de valeurs plus générales que celles attendues est également considérée – n_{gen}. Et enfin le taux d'erreur est indiqué (valeurs proposées qui sont totalement différentes et décorrélées de la valeur attendue) – n_{faux}. Notons que le modèle M_1 correspondant à la

méthode *Sums* est équivalent au modèle M_2 pour lequel la valeur de γ est égale à zéro (première ligne de M_2). De même, le modèle M_3 est équivalent au modèle M_4 pour lequel $\gamma = 0$. En effet, dans ce cas les règles ne sont pas prises en compte dans le calcul. Nous avons déjà montré dans (Beretta et al., 2016) que la prise en compte de la propagation d'information en fonction de l'ontologie du domaine (M_3) apportait une plus-value par rapport à l'approche classique (M_1).

A l'inverse, considérer uniquement l'influence des règles lors du processus de recherche de vérité consisterait à choisir comme valeur $\gamma = 1$. Dans la grande majorité des cas, cette configuration offre les résultats les moins bons si l'on considère le nombre de valeurs vraies attendues et le taux d'erreur. Ce résultat s'explique facilement. Les règles reposent uniquement sur une analyse statistique de KB et peuvent parfois ne pas être valides pour toutes les entités. Par contre, cette configuration fournit toujours le meilleur taux en termes de valeur générique. Ce constat confirme l'intuition suivante : l'application des règles tend à favoriser une connaissance générique. En effet, plus le recouvrement de la tête d'une règle est élevé, plus le nombre d'instances pour lequel elle est valide est grand. Pour les valeurs plus génériques, il est donc d'autant plus facile de trouver des règles qui seront vérifiées.

Cependant, même si l'application des règles d'association favorise les valeurs plus génériques lors de la recherche de vérité, il est intéressant de les prendre en compte. Elles permettent en effet, en accordant plus de confiance dans les valeurs génériques, de favoriser certaines branches lors de l'exploration de l'arbre des valeurs par l'algorithme glouton de sélection des valeurs vraies. Elles permettent donc d'éviter certaines erreurs lors de l'amorce du processus de sélection des valeurs vraies.

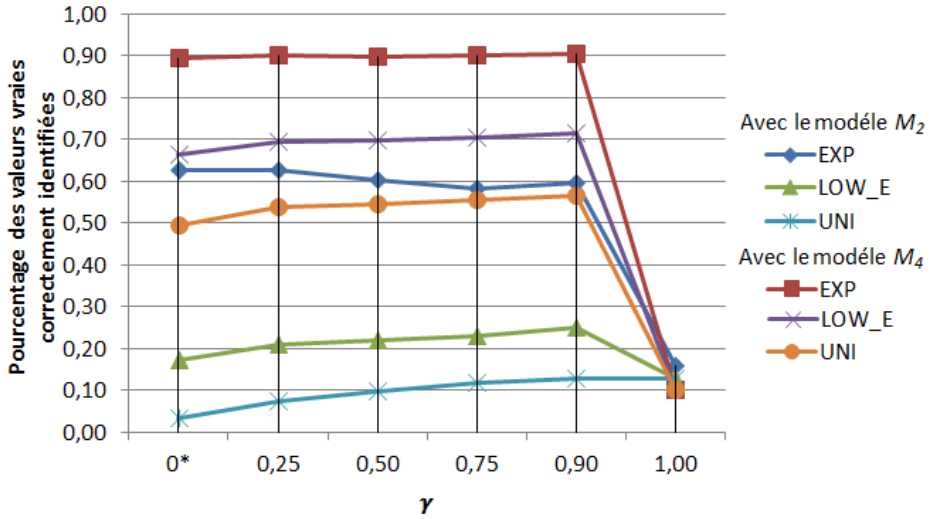


FIGURE 2 – Synthèse des résultats : précision (taux de valeurs attendues obtenues) en fonction du jeu de test, de la méthode utilisée et de γ.

Si l'on considère maintenant le modèle M_2 (*Sums+Règles*), on remarque une amélioration des résultats par rapport à la méthode de base, même si cette amélioration est moins flagrante qu'avec les autres modèles (M_3 – *Sums adapté* et M_4 – *Sums adapté+règles*). Cette amélioration s'explique par la propagation des confiances sur les valeurs (en respectant l'ordre partiel donné par l'ontologie de domaine). Cette propagation dans le cas des modèles M_3 et M_4 concerne l'ensemble des valeurs alors que les règles ne concernent, quant à elles, qu'une sous partie de ces valeurs.

Tous modèles confondus, les meilleurs résultats sont obtenus avec la méthode M_4 et un coefficient $\gamma = 0.9$, et ce sur tous les types de corpus testés. Cette configuration permet d'obtenir le plus grand nombre de valeurs vraies attendues, et de diminuer le taux d'erreur. Sur ces deux critères, ce gain est d'autant plus grand que la disparité (contradictions) entre les sources est grande.

L'importance de la valeur de γ est également à souligner. On le voit bien sur les résultats obtenus avec le modèle M_2 (pour lequel on ne tient pas compte de la propagation sur les valeurs proposées). Pour le jeu de données EXP dans lequel les sources ont tendance à être plus en accord sur la valeur proposée et cette valeur étant très spécifique, les meilleurs résultats obtenus en terme de précision sont avec un coefficient $\gamma = 0.25$. On voit donc que si l'on est dans un cas où les sources sont relativement fiables sur un sujet donnés (site Web spécialisé, par exemple), il est préférable d'accorder plus de confiance aux déclarations qu'elles proclament. Par contre, pour les deux autres types de jeu de données (LOW_E et UNI) dans lesquels les désaccords entre les sources sont nombreux et vont en croissant, il est préférable de s'appuyer sur les règles d'association identifiées ($\gamma = 0.9$ dans le premier cas et $\gamma = 1$ dans le cas de UNI). La comparaison des performances obtenues au travers des différents modèles est illustrée dans la **FIGURE 2**.

5 Conclusion et perspectives

A l'heure où la détection de vérité devient de plus en plus cruciale pour nombre d'applications, il nous semble indispensable de développer des approches de recherche de vérité qui tiennent compte d'une modélisation de connaissance sous forme d'ontologies. Dans une contribution précédente, nous avons montré comment certaines relations de cette ontologie permettait d'améliorer les approches existantes. Dans cette présente contribution, nous montrons que la A-Box associée a également une grande influence et peut améliorer le processus de recherche de vérité. En considérant l'ordre partiel qui existe entre les valeurs proposées par différentes sources, l'utilisation de règles d'association collectées après l'analyse de la A-Box, permet de favoriser certaines valeurs plus génériques et ainsi d'améliorer la stratégie de sélection des valeurs vraies. En fonction du contexte, une bonne paramétrisation permettra d'obtenir de meilleurs résultats que les approches classiques. Nous souhaitons compléter cette étude en considérant d'autres méthodes de référence comme *AverageLog*, *Investment* et *PooledInvestment* (Pasternack & Roth, 2010), et *Cosine* et *2-Estimated* (Galland *et al.*, 2010). Ces développements sont en cours. Nous souhaitons également effectuer des tests sur d'autres jeux de données avec des prédicats plus ou moins spécialisés par rapport à un domaine. La propagation vers les concepts plus généraux (selon la propagation des *croyances*) produit des améliorations. Il est aussi souhaitable d'effectuer une propagation vers les concepts plus spécifiques suivant le mode de propagation des *possibilités* en théorie des croyances (modèle en cours de définition). Enfin, nous souhaitons également intégrer ces modules logiciels dans une chaîne réelle d'enrichissement de bases de connaissances basée sur une extraction à partir de textes.

Références

AUER, S., BIZER, C., KOBILAROV, G., LEHMANN, J., CYGANIAK, R., & IVES, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. In K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L.

Nixon, … P. Cudré-Mauroux (Eds.), *The Semantic Web, Lecture Note in Computer Science* (Vol. 4825, pp. 722–735). Springer Berlin Heidelberg.

BERETTA, V., HARISPE, S., RANWEZ, S., & MOUGENOT, I. (2016). Utilisation d'ontologies pour la quête de vérité : une étude expérimentale. In *Actes des 27es Journées francophones d'Ingénierie des Connaissances IC2016*. Montpellier, France.

BERTI-ÉQUILLE, L., & BORGE-HOLTHOEFER, J. (2015). *Veracity of Data: From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics* (Synthesis). Morgan & Claypool Publishers.

BLANCO, L., CRESCENZI, V., MERIALDO, P., & PAPOTTI, P. (2010). Probabilistic models to reconcile complex data from inaccurate data sources. In B. Pernici (Ed.), *Advanced Information Systems Engineering: 22nd International Conference CAiSE 2010 Proceedings* (pp. 83–97). Hammamet, Tunisia: Springer-Verlag.

DONG, X. L., BERTI-EQUILLE, L., HU, Y., & SRIVASTAVA, D. (2010). Global detection of complex copying relationships between sources. In E. Bertino, P. Atzeni, K. L. Tan, Y. Chen, & Y. C. Tay (Eds.), *Proceedings of the VLDB Endowment* (Vol. 3, pp. 1358–1369). VLDB Endowment.

FENO, D. R. (2007). *Mesures de qualité des règles d'association : normalisation et caractérisation des bases*. Université de la Réunion, France.

GALARRAGA, L., TEFLIOUDI, C., HOSE, K., & SUCHANEK, F. M. (2015). Fast rule mining in ontological knowledge bases with AMIE. *The VLDB Journal The International Journal on Very Large Data Bases*, *24*(6), 707–730.

GALLAND, A., ABITEBOUL, S., MARIAN, A., & SENELLART, P. (2010). Corroborating information from disagreeing views. In *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10* (pp. 131–140). New York, New York, USA: ACM Press.

JEAN, P.-A., HARISPE, S., RANWEZ, S., BELLOT, P., & MONTMAIN, J. (2016). Uncertainty Detection in Natural Language: A Probabilistic Model. In *6th International Conference on Web Intelligence, Mining and Semantics, WIMS'16* (p. 10:1-10:10). Nîmes, France: ACM Press, New York (USA).

LI, Y., GAO, J., MENG, C., LI, Q., SU, L., ZHAO, B., … HAN, J. (2015). A Survey on Truth Discovery. *ACM SIGKDD Explorations Newsletter*, *17*(2), 1–16.

MAIMON, O., & ROKACH, L. (2005). *Data Mining and Knowledge Discovery Handbook*. (O. Maimon & L. Rokach, Eds.). Springer US.

MENG, C., JIANG, W., LI, Y., GAO, J., SU, L., DING, H., & CHENG, Y. (2015). Truth Discovery on Crowd Sensing of Correlated Entities. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems, SenSys '15* (pp. 169–182). Seoul, South Korea: ACM Press, New York (USA).

PASTERNACK, J., & ROTH, D. (2010). Knowing What to Believe (when you already know something). In *23rd International Conference on Computational Linguistics, COLING'10* (pp. 877–885). Stroudsburg, PA, USA: Association for Computational Linguistics.

POCHAMPALLY, R., DAS SARMA, A., DONG, X. L., MELIOU, A., & SRIVASTAVA, D. (2014). Fusing data with correlations. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data - SIGMOD '14* (pp. 433–444). New York, New York, USA: ACM Press.

QI, G.-J., AGGARWAL, C. C., HAN, J., & HUANG, T. (2013). Mining collective intelligence in diverse groups. In *Proceedings of the 22nd international conference on World Wide Web - WWW '13* (pp. 1041–1052). New York, New York, USA: ACM Press.

QUBOA, Q., & SARAEE, M. (2013). A state-of-the-art survey on semantic web mining. *Intelligent Information Management*, *5*(1), 10–17.

WANG, D., ABDELZAHER, T., & KAPLAN, L. (2015). *Social Sensing: Building Reliable Systems on Unreliable Data*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

WANG, S., SU, L., LI, S., HU, S., AMIN, T., WANG, H., … ABDELZAHER, T. (2015). Scalable Social Sensing of Interdependent Phenomena. In *Proceedings of the 14th International Conference on Information Processing in Sensor Networks, IPSN '15* (pp. 202–213). Seattle, Washington: ACM, New York, NY, USA.

WANG, Z., & LI, J. (2015). RDF2Rules: learning rules from RDF knowledge bases by mining frequent predicate cycles. *arXiv Preprint arXiv:1512.07734*.

YIN, X., HAN, J., & YU, P. S. (2008). Truth Discovery with Multiple Conflicting Information Providers on the Web. *IEEE Transactions on Knowledge and Data Engineering*, *20*(6), 796–808.

Génération automatique d'un questionnaire à partir d'une ontologie de domaine

Leila Zemmouchi-Ghomari[1], Faiza Deghmani[2] et Aya Meghnous[3]

[1,2,3] Département d'informatique, Faculté d'Electronique et d'Informatique,
Université des Sciences et des Technologies Houari Boumediene USTHB
Lzemmouchi-ghomari@usthb.dz
{deghmani.faiza, meghnous.aya}@gmail.com

Résumé : L'usage croissant des ontologies dans les applications informatiques pose le problème de l'estimation de la qualité de ces ontologies et de leur adéquation par rapport au problème qu'elles contribuent à résoudre. La validation d'une ontologie vérifie que la sémantique des définitions modélise la réalité pour laquelle l'ontologie a été développée. Etant donné que le langage naturel est considéré comme un moyen incontournable de transfert de la connaissance et que les experts du domaine ne maîtrisent pas forcément les langages formels des ontologies, une des techniques de validation les plus utilisées est le questionnaire en langage naturel. Dans cet article, nous proposons un système de génération automatique de questionnaire à partir d'ontologies dans le cadre d'une approche d'évaluation des ontologies proposée dans un travail antérieur. Le résultat de ce système est un questionnaire destiné à représenter les composants de l'ontologie de façon informelle afin d'être évalués par les experts du domaine.

Mots-clés : Ontologies de domaine, validation d'ontologies, évaluation des ontologies, questionnaire, concepteur d'ontologies, expert du domaine.

1 Introduction

Le mouvement mondial du développement de technologies open source, libres et permettant le partage et la réutilisation ne cesse de progresser. Parmi ces technologies, les ontologies, « spécifications explicites de conceptualisation » (Gruber, 1993), occupent une place de choix dans la palette des technologies du web Sémantique. En effet, les ontologies jouent un rôle primordial dans la représentation, la réutilisation et le partage des connaissances d'un domaine de façon formelle, consensuelle et explicite.

Néanmoins, l'usage croissant d'ontologies formelles dans les applications informatiques pose le problème de l'évaluation de la qualité de ces ontologies et de leur adéquation aux problèmes qu'elles contribuent à résoudre. L'évaluation des ontologies selon Gomez Perez (Gomez-Perez, 2004a), porte sur un jugement technique du contenu de l'ontologie au regard du cadre de référence (spécification des besoins) et s'assure que l'ontologie modélise correctement le domaine du monde réel pour lequel elle a été développée.

De façon générale, l'évaluation des ontologies est incontournable dans le cas du développement d'une nouvelle ontologie ou dans le cas de la réutilisation d'une ontologie existante (Sheth et al., 2010). En effet, l'évaluation des ontologies est une phase critique dans le domaine de l'ingénierie ontologique (Eschenbach & Gruninger, 2008). Plusieurs méthodologies de développement d'ontologies (Fernandez-Lopez et al., 1997), (Sure & Studer, 2002), (Suárez –Figueroa et al., 2012) incluent l'évaluation comme partie intégrante du processus de développement des ontologies (Lovrenčić & Čubrilo, 2008). En effet, les ontologies de qualité incertaine posent plusieurs problèmes. Les agents qui exploitent ces

ontologies et qui utilisent des connaissances incomplètes, inconsistantes, non pertinentes ou ambigües auront des difficultés à obtenir des résultats probants (Burton-Jones et al., 2005). Une ontologie jugée de bonne qualité garantira l'absence d'erreurs, et une réutilisation au moindre risque (Gomez-Perez, 2004b).

Gomez-Perez dans (Gomez-Perez, 2004c) distingue deux types d'évaluation des ontologies: la vérification et la validation des ontologies. La vérification consiste à s'assurer que l'ontologie a été construite correctement alors que la validation mesure le degré d'adéquation entre l'ontologie et le domaine du monde réel qu'elle représente. C'est ce dernier type d'évaluation qui nous intéresse dans le cadre du présent travail.

Les approches de validation des ontologies peuvent être classées selon différentes dimensions (Brank et al., 2005), à savoir :

- Comparaison avec une référence ou un « gold standard », par exemple une ontologie de haut niveau, générique ou une ontologie de référence (Zemmouchi-Ghomari & Ghomari, 2009).
- Comparaison avec une source de données, par exemple, une collection de documents qui couvrent le domaine.
- Evaluation par les humains, à savoir : les développeurs d'ontologie, les experts du domaine et les utilisateurs finaux.
- Evaluation basée sur une application intégrant une ontologie, en d'autres termes, vérifier si l'application répond bien aux spécifications initiales.

Au-delà de cette classification, toutes ces approches nécessitent la coopération des experts du domaine. D'ailleurs, l'évaluation des ontologies est souvent assimilée à un processus semi-automatique (Gomez-Perez, 2004c) prenant en considération l'intervention humaine. De plus, l'ontologie est un artefact social puisqu'elle représente un consensus d'une conceptualisation partagée par un ensemble de parties prenantes. Le consensus est mesuré à travers la proportion d'accord que les experts du domaine partagent concernant les éléments constituants de l'ontologie (Gangemi et al., 2005).

L'objectif du présent travail est de proposer un outil de génération automatique de questionnaire à partir d'une ontologie de domaine dans le cadre d'une approche d'évaluation d'ontologie proposée dans un travail antérieur (Zemmouchi-Ghomari & Ghomari, 2014).

Cet article est organisé de la manière suivante: Section 2 présente l'approche d'évaluation proposée qui constitue le cadre du présent travail. Section 3 décrit la conception, l'implémentation et l'expérimentation du système automatique de génération d'un questionnaire à partir d'une ontologie. Section 4 décrit les travaux similaires, à savoir : l'évaluation des ontologies par les experts du domaine à travers un questionnaire dans la littérature. Section 5 conclut cet article et présente quelques extensions possibles pour améliorer l'outil proposé.

2 Approche proposée

Dans un travail précédent (Zemmouchi-Ghomari & Ghomari, 2014), nous avons proposé une approche d'évaluation qui débute à partir d'une ontologie à évaluer et qui s'achève avec une ontologie mise à jour selon les recommandations des évaluateurs (voir Fig. 1). Cette approche est composée de 5 étapes :

1. Dérivation d'un questionnaire à partir des composants de l'ontologie.
2. Agrégation des résultats de l'enquête auprès des experts.
3. Analyse et synthèse des résultats obtenus.
4. Mise à jour du questionnaire en fonction du feedback des experts.
5. Mise à jour de l'ontologie en fonction des connaissances issues des résultats.

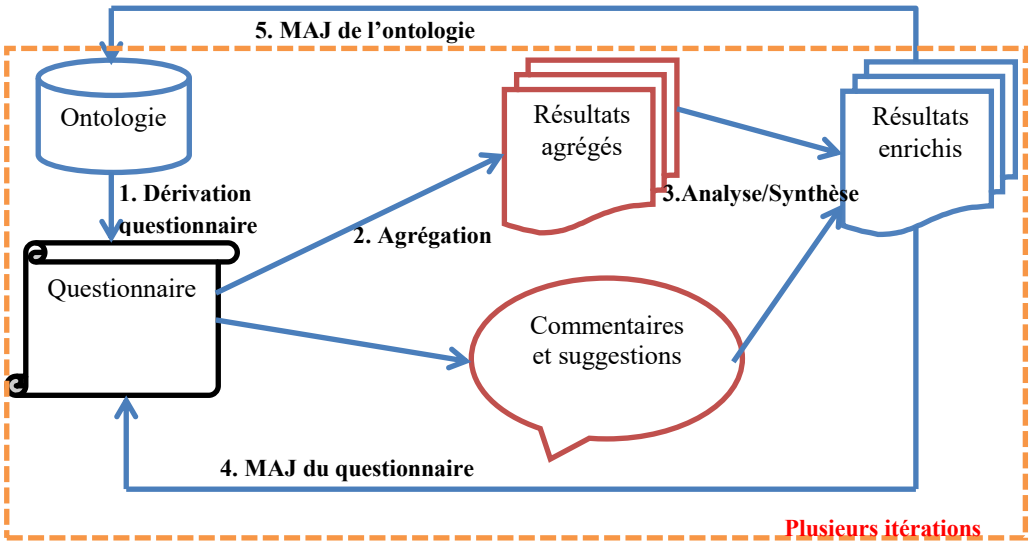


FIG 1. Approche de validation d'une ontologie par les experts via un questionnaire.

Dans cet article, nous allons nous concentrer sur les étapes 1 et 2 de l'approche proposée. Nous nous contenterons de fournir quelques suggestions concernant le traitement du feedback du questionnaire dans les étapes 3 et 4.

Les recommandations des experts sont transmises au concepteur de l'ontologie en vue de de procéder à sa mise à jour (étape 5).

Etape 1 et 2: Dérivation du questionnaire à partir des composants de l'ontologie et génération de résultats agrégés

Le questionnaire est un outil de collection et d'enregistrement d'informations sur un sujet d'intérêt. Il est composé d'une liste de questions ainsi que d'un espace dédié aux réponses. Il doit avoir un objectif préalablement défini.

Le questionnaire est destiné à représenter les composants de l'ontologie de façon informelle (langage naturel) afin d'être évalués par les experts du domaine qui ne sont pas nécessairement connaisseurs en matière de langages de représentation des ontologies, tels que : RDF[1] (Resource Description Framework) ou OWL[2] (Web Ontology Language).

Une ontologie est évaluée sur la base de ses composants. Des classes organisées hiérarchiquement, reliées par des propriétés d'objet et décrites par des propriétés de données. L'ontologie est interprétée grâce aux axiomes, peuplée par des instances et documentée via des annotations.

De ce fait, l'approche est conçue de telle sorte à décomposer l'ontologie dans son ensemble en différents niveaux, à savoir: (a) hiérarchie des classes, (b) axiomes, (c) propriétés d'objet (d) propriété de données (e) instances et (f) annotations.

Afin de pallier à la génération manuelle du questionnaire à partir de l'ontologie (surtout pour les ontologies de grande taille), nous décrivons dans la section 3 : la conception et l'implémentation d'un outil qui prend en charge la réalisation automatique des étapes 1 et 2 de l'approche proposée.

Etape 3 et 4 : Analyse/Synthèse des résultats obtenus et Mise à jour du questionnaire

Holsapple et Joshi dans (Holsapple & Joshi, 2005) ont proposé une méthode d'évaluation pour la collaboration manuelle dans le cadre de l'ingénierie ontologique, dans laquelle chaque

[1] https://www.w3.org/TR/rdf-primer/

[2] https://www.w3.org/TR/owl-features/

suggestion faite par un expert est évaluée par d'autres experts, de ce fait, les meilleures suggestions sont celles qui obtiennent l'approbation de plusieurs experts.

Une technique connue pour appliquer ce principe est la méthode Delphi, développée par Dalkey et Helmer en 1963 (Dalkey & Helmer, 1963). C'est une méthode largement utilisée et acceptée pour obtenir la convergence des opinions des experts concernant une thématique particulière (Hsu & Sandford, 2007).

Le processus Delphi est exécuté en plusieurs itérations jusqu'à ce qu'un consensus soit atteint. L'expérience a montré que trois itérations sont suffisantes pour atteindre ce consensus dans la plupart des cas.

Nous pensons que cette méthode est appropriée pour traiter le feedback des experts sur le questionnaire jusqu'à ce qu'un consensus soit atteint. Selon (Gomez-Perez, 2004c), il est préférable que le groupe d'experts soit restreint en nombre à condition qu'il soit composé d'experts reconnus dans le domaine de l'ontologie à évaluer. La mise à jour du questionnaire est réalisée sur la base des commentaires des experts qui ne seraient pas d'accord avec la formulation des questions générées et proposeraient soit de les supprimer soit de les modifier.

3 Conception du système automatique de génération de questionnaire à partir d'une ontologie

L'objectif de cet outil est de permettre aux concepteurs d'ontologies de soumettre leurs ontologies aux experts du domaine à travers des questionnaires générés automatiquement à partir de ces ontologies.

3.1. Fonctionnalités du système

Nous allons définir dans ce qui suit les fonctionnalités du système à travers deux acteurs possibles, le concepteur d'ontologie et l'expert du domaine.

1. Concepteur d'ontologie : Son rôle est de tester le questionnaire généré, et de vérifier si les questions correspondent aux composants de l'ontologie qu'il a conçue. Le rôle du concepteur, est résumé comme suit :

- Le concepteur vérifie l'adéquation entre son ontologie source et les questions générées.

Cette adéquation peut être totale, partielle ou nulle selon le degré d'adéquation du questionnaire généré par rapport à l'ontologie source.

Le concepteur retourne son feedback concernant chaque question générée, pour une éventuelle mise à jour de l'outil de génération du questionnaire et par conséquent la modification du questionnaire lui-même.

Lors d'une étape ultérieure, Le concepteur mettra à jour son ontologie dans le cadre de la validation de l'ontologie via le questionnaire, en fonction des réponses des experts de domaine au questionnaire.

2. Expert du domaine : son rôle est de répondre au questionnaire, après la validation de ce dernier par le concepteur d'ontologie. Les réponses proposées obéissent aux critères de qualité des ontologies énoncés dans la section 3.4. Ces réponses seront analysées par la suite pour la mise à jour de l'ontologie. L'expert a également la possibilité de critiquer le questionnaire sous forme de texte libre, ces remarques sont prises en considération pour la mise à jour de ce dernier.

3.2. Modélisation des données du système

Dans le cadre de ce travail, nous nous focalisons, dans un premier temps, sur la version 1 du langage OWL DL comme langage de représentation de connaissances. Ce choix se justifie par le fait qu'OWL soit considéré comme le langage le plus expressif des langages

ontologiques. Il constitue une extension à RDF et RDFS[3] puisqu'il intègre tous leurs constructeurs et offre également un niveau d'expressivité plus élevé en matière de spécification des classes, telles que: l'équivalence entre classes, l'incompatibilité ainsi que la symétrie et la transitivité des propriétés à titre d'exemple.

Les règles de gestion sont les suivantes : un questionnaire comporte plusieurs types de questions, chaque question est générée selon un canevas ou patron prédéfini, qui fait référence à un seul composant ontologique.

Ces questions doivent être testées et analysées par le concepteur de l'ontologie. Le feedback du concepteur sur le questionnaire généré est stocké sous forme d'appréciations sur les questions que nous appellerons « Review Questions » qui peuvent être positives ou négatives. Dans le cas où l'analyse du questionnaire est négative, ce dernier est modifié, si nécessaire en plusieurs itérations. Une fois le questionnaire stabilisé par le concepteur de l'ontologie, les experts du domaine répondent au questionnaire. Ces réponses sont stockées puis agrégées et analysées pour la mise à jour de l'ontologie (voir Fig. 2).

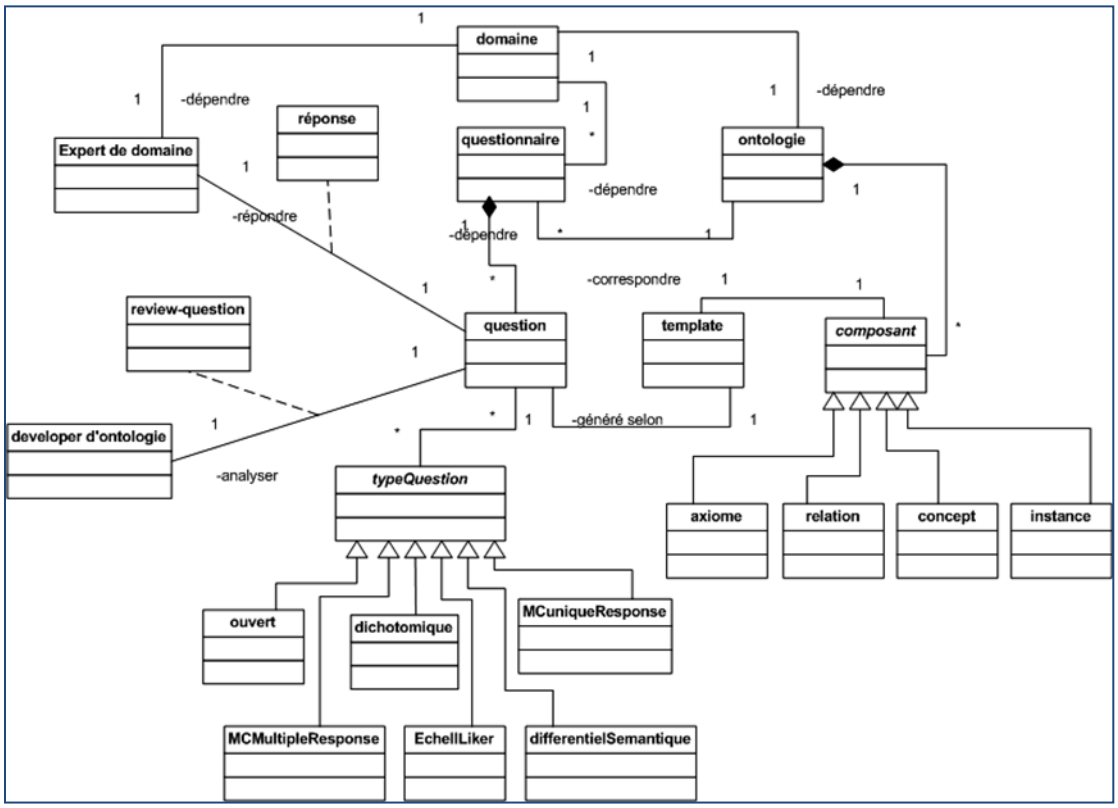


FIG 2. Méta Modèle du questionnaire d'évaluation d'une ontologie.

3.3. Conception des patrons de questions relatifs aux composants de l'ontologie

Chaque élément ontologique génère une question. Nous définissons une question générique sous forme d'une expression régulière contenant une partie fixe (en minuscule), et une partie variable (en majuscule et en rouge) pour chaque élément ontologique.

Nous présentons dans ce qui suit (table 1) quelques patrons[4] utilisés pour représenter les composants d'une ontologie:

[3] https://www.w3.org/TR/rdf-schema/

[4] La totalité des patrons est disponible en téléchargement à cette adresse :

TABLE 1 – *Quelques patrons des composants de l'ontologie en OWL 1 DL.*

Elément ontologique	Question générique ou patron	Exemple associé
Sous-Classes OWL:subClalssOf	Every CLASS1 is a/an CLASS2	Every MAN is a/an PERSON
	Every CLASS1 is a/an CLASS2 AND/OR CLASS3 AND/OR….. CLASSn	Every PARENT is a/an MOTHER OR a/an FATHER Every MOTHER is a/an PARENT AND a/an WOMAN
Restrictions de quantification (RQ) Owl:AllValuesFrom	All PROPERTY of a/an CLASS1 are CLASS2	All hasCreator of a/an BOOK are AUTHOR
Restrictions de cardinalité (RC) Owl :MinCardinality	CLASS on PROPERTY has at least CARDIN-MIN	PARENT on hasChild has at least 1
Classes d'Individus (CI)	OWL:one of Member(s) of CLASS is/are INSTANCE1, INSTANCE2…INSTANCEn	Member(s) of DaysOfWeek is/are Saturday, Sunday…Friday
Jointure des Classes (JC)	OWL:IntersectionOf CLASS1 and CLASS2 and …CLASSn	A MOTHER is FEMALE and PARENT
Classes Equivalentes (CE)	OWL: EqivalentClass CLASS1 is equivalent to CLASS2	A PERSON is equivalent to HUMAN
Classes Disjointes (CD)	OWL: DisjointWith CLASS1 is not CLASS2	A MAN is not a WOMAN
Propriétés objets ou Propriétés de types de données	OWL:ObjectProperty, OWL:DataTypeProperty rdfs:domain rdfs:range CLASS1 ObjectProperty CLASS2/PropertyValue	TEACHER GivesGradeTo STUDENT
Instances	rdf:type INSTANCE is a/an instance of CLASS	MARIA is a/an instance of WOMAN
Annotations	rdfs:Comment RESSOURCE has comment TEXT	DBPEDIA has comment "an ontology based on wikipedia"

3.4. Mise en correspondance entre les critères de qualité des ontologies et les réponses possibles dans le questionnaire

Afin de proposer les réponses possibles aux questions incluses dans le questionnaire destiné aux experts de domaine, nous nous sommes basés sur quelques critères de qualité des ontologies tels qu'ils ont été définis par (Gruber, 1995), (Fox et al., 1996), (Gomez-Perez, 2004c) et (Obrst et al., 2007). Les réponses sont récapitulées dans le tableau suivant:

TABLE 2 – *Les critères de qualité des ontologies associés aux réponses des questions.*

Critères de qualité	Réponses possibles	Elément ontologique
Clarté	Clair/Pas clair	Classes, axiomes, propriétés d'objets, propriétés de données et annotations

https://sourceforge.net/projects/ontologyquestvalidator/files/OntologyQuestionnaireApp/Component%20patterns%20of%20an%20OWL%201%20DL%20Ontology.pdf/download

Cohérence	Cohérent/Contradictoire	Annotations
Précision	Vrai/Faux Toujours/Parfois/Jamais	Classes Axiomes
Qualité Syntaxique	Vrai/vrai mais un autre terme serait plus approprié/faux	propriétés d'objets, propriétés de données et axiomes
Qualité pragmatique	Pertinent/Non Pertinent	Propriétés de données, annotations

1 La Clarté de tous les composants ontologiques est un critère important pour le partage des connaissances entre les différents acteurs appartenant à une communauté.

2 La Cohérence des annotations pour vérifier si les descriptions des éléments ontologiques ne conduisent pas à des contradictions par rapport aux définitions des éléments ontologiques.

3 La Précision afin d'éviter toute ambiguïté. La précision est plus subtile lors de la validation des énoncés des axiomes, c'est pourquoi nous avons proposé plus de deux options booléennes.

4 La Qualité syntaxique des propriétés de données, des propriétés d'objet et des axiomes. Les expressions qui en résultent doivent être syntaxiquement correctes.

5 La Qualité pragmatique ou Pertinence des propriétés de données et des annotations puisqu'elles ont été définies pour décrire d'autres composants tels que les classes. La pertinence des propriétés de données proposées doit être vérifiée, car la plupart d'entre elles sont proposées par le concepteur de l'ontologie contrairement aux autres éléments ontologiques (classes, propriétés d'objet) qui sont dérivées des spécifications de l'ontologie (par exemple via les questions de compétence). Par exemple: la date et le lieu de naissance; propriétés de la classe Étudiant ne sont probablement pas dérivés de la phase de spécification de l'ontologie, bien qu'ils soient souvent prévus dans la liste des propriétés de la classe Etudiant.

3.5. Calcul du taux d'adéquation ontologie-questionnaire

Tel que précisé dans les spécifications du système (3.1), nous commençons par soumettre le questionnaire généré au concepteur de l'ontologie. Ce dernier évalue chaque question par rapport à l'élément ontologique correspondant. Pour cela nous lui proposons trois possibilités : adéquation totale (T), adéquation partielle (P) et adéquation nulle (N).

Pour calculer le pourcentage d'adéquation ontologie-questionnaire, nous avons procédé comme suit, pour chaque élément ontologique (classe, axiome, propriété d'objet, propriété de donnée, instance et annotation), nous avons calculé le pourcentage de son adéquation avec les questions générées en fonction des réponses du concepteur de l'ontologie sur le questionnaire.

Dans ce qui suit, nous présentons un extrait de l'algorithme de calcul correspondant à l'élément ontologique Classe à titre d'exemple:

```
if (Type(Elément) == Classe) then
   TClasse = (Pourcentage de réponses « adéquation totale ») *
   Facteur d'importance des classes
   PClasse = (Pourcentage de réponses « adéquation partielle ») *
   Facteur d'importance des classes
   NClasse = (Pourcentage de réponses « adéquation nulle ») *
   Facteur d'importance des classes
endif
```

Où le facteur d'importance représente pour chaque composant son degré d'importance dans une ontologie. Ce facteur est déterminé par le concepteur de l'ontologie. Par exemple, les propriétés de données et les propriétés d'objets sont très importantes dans une ontologie comme FOAF, dans une autre ontologie comme DBpedia les instances auront une importance

capitale. Si le concepteur estime que tous les éléments ontologiques ont la même importance, le facteur sera égal à 1.

$T = T_{Classe} + T_{Axiome} + T_{Propriété\ d'objet} + T_{Propriété\ de\ donnée} + T_{Instance} + T_{Annotation}$
$P = P_{Classe} + P_{Axiome} + P_{Propriété\ d'objet} + P_{Propriété\ de\ donnée} + P_{Instance} + P_{Annotation}$
$N = N_{Classe} + N_{Axiome} + N_{Propriété\ d'objet} + N_{Propriété\ de\ donnée} + N_{Instance} + T_{Annotation}$

Et pour calculer le pourcentage global de l'adéquation ontologie-questionnaire, nous avons défini une pondération en fonction des réponses du concepteur.

$Adéquation\ pondérée = T+P+N$
$T_{Pondéré} = T/Adéquation\ pondérée.$
$P_{Pondéré} = P/Adéquation\ pondérée.$
$N_{Pondéré} = N/Adéquation\ pondérée.$

Ces statistiques constituent la concrétisation de l'étape 2 de l'approche, à savoir l'agrégation des résultats du questionnaire.

3.6. Implémentation du système

L'implémentation a été réalisée avec J2EE en utilisant les technologies JSP (Java Servlet Pages), le serveur web: Apache Tomcat, le SGBD: PostGreSQL et l'API Jena pour la manipulation des documents RDF et OWL. Notre application est disponible en téléchargement à l'adresse suivante : https://sourceforge.net/projects/ontologyquestvalidator/.

Deux menus sont proposés en page d'accueil: un menu concepteur de l'ontologie et un menu expert du domaine.

L'ontologie est identifiée par son URL, chemin vers le disque local ou avec son code source. Dans chaque menu le même questionnaire est généré avec des réponses différentes, puisque le concepteur jugera du degré d'adéquation de la question avec l'élément ontologique correspondant (Fig. 3) et l'expert du degré d'adéquation de la question avec le domaine couvert par l'ontologie.

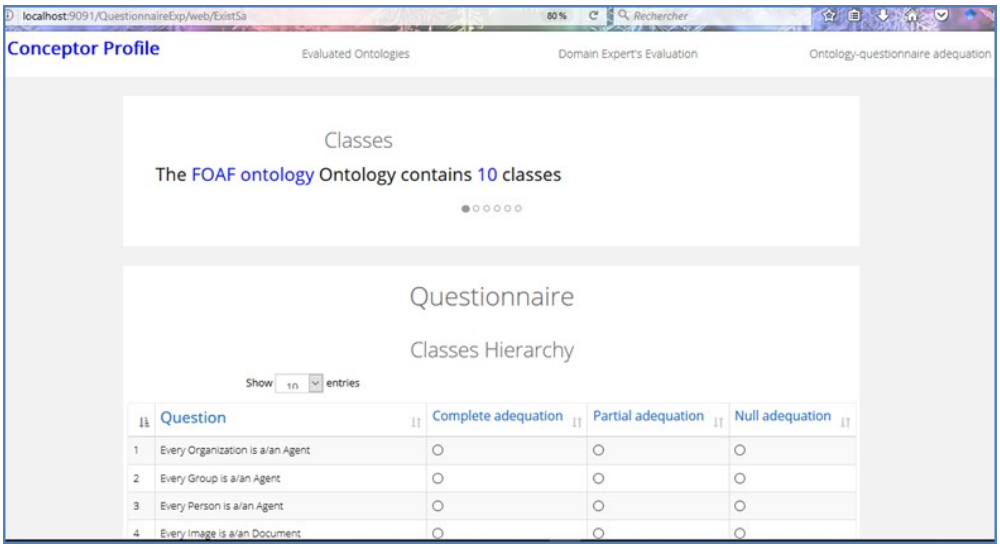


FIG 3. Questionnaire de l'ontologie FOAF (Partie de la hiérarchie des classes)

Des statistiques par élément ontologique sont générées pour afficher les résultats des différentes évaluations (des captures d'écran sont disponibles sur le site de téléchargement de l'outil).

3.7. Expérimentation de l'outil de génération de questionnaire

Nous avons soumis notre application à trois concepteurs d'ontologies qui ont accepté d'évaluer la qualité du questionnaire généré et son degré d'adéquation avec l'ontologie à l'origine (voir table 3).

TABLE 3 – *Résultats de l'expérimentation de l'outil de génération de questionnaire par des développeurs d'ontologies.*

Nom / URL de l'ontologie	**Description**	**Elément de l'ontologie/ Pondération**	**Adéquation Totale**	**Adéquation Partielle**	**Adéquation Nulle**
Ontologie de compétences **ECAO**[5]**,** Esi Competency Application Ontology (67 classes, 12 associations, 41 attributs, 49 instances)	Cette ontologie vise à représenter les compétences au sein d'une école d'informatique (ESI) dans le cadre de son **système de localisation de compétences**. (Langue: Français)	Classes / 2	100%	0%	0%
		Axiomes / 1	62.5%	0%	37.5%
		Propriétés d'objets / 1	60%	40%	0%
		Instances / 2	100%	0%	0%
		Adéquation Totale globale : 87.08% Adéquation Partielle globale : 6.67% Adéquation Nulle globale : 6.25%			
Ontologie **Annotation** (219 classes, 31 associations, 8 attributs, 21 instances)	Cette ontologie vise à formaliser **les objectifs d'annotations subjectifs de l'enseignant** sur les documents de travail lors de la réalisation de ses activités pédagogiques (Langue: Français)	Classes / 4	100%	0%	0%
		Axiomes / 1	90.48%	0%	9.52%
		Propriétés d'objets / 3	33.33%	0%	66.67%
		Instances / 2	100%	0%	0%
		Adéquation Totale globale : 79.05% Adéquation Partielle globale : 0% Adéquation Nulle globale : 20.95%			
Cook Ontology (25 classes, 8 associations, 7 attributs, 4 instances)	Cette ontologie a été conçue pour étendre la **description des produits alimentaires** qu'offre l'ontologie Food (http://data.lirmm.fr/ontologies/food) (Langue: Anglais)	Classes / 1	100%	0%	0%
		Axiomes / 1	91.67%	0%	8.33%
		Propriétés d'objets / 1	77.78%	0%	22.22%
		Instances / 1	100%	0%	0%
		Adéquation Totale globale : 92.36% Adéquation Partielle globale : 0% Adéquation Nulle globale : 7.64%			

Les résultats montrent une adéquation totale globale supérieure à 79% pour les ontologies testées. Toutefois, les développeurs d'ontologies ont soulevé le problème de la non prise en charge d'autres langues d'ontologies que la langue Anglaise. En effet, étant donné que les patrons des questions sont en Anglais, les questions générées sont ambigües.

Une autre limitation de l'outil concerne le langage des questions qui n'est pas naturel à 100%, ce qui constitue un obstacle pour l'évaluation des experts, voici un exemple d'une question générée: *PartieTangible concerne_tangiblement_une_annotation Annotation*.

De plus, le questionnaire généré est long, étant donné les composants variés des ontologies, ce qui peut constituer un frein pour répondre à l'ensemble des questions.

Nous avons également constaté que les développeurs ont tendance à évaluer l'adéquation ontologie/domaine de connaissances comme des experts du domaine, au lieu d'évaluer l'adéquation ontologie/questionnaire.

[5] *https://sourceforge.net/projects/competenyapplicationontology/*

L'évaluation de l'outil de génération de questionnaire devra se poursuivre avec des ontologies variées en termes de taille, de langage de représentation de connaissances, de domaine et de langue.

Lorsque l'outil sera jugé satisfaisant au regard d'un nombre significatif de concepteurs d'ontologies, son exploitation effective sera envisageable. En d'autres termes, il pourra être utilisé pour évaluer l'adéquation ontologie-domaine de connaissances, cette fois-ci par les experts du domaine.

4 Travaux similaires

Pour valider les ontologies suivant l'approche basée sur l'évaluation humaine, plusieurs travaux ont été proposés, nous nous intéressons aux travaux basés sur Les questionnaires d'évaluation :

Dans (Papasalouros et al., 2008), la génération automatique d'un questionnaire à choix multiple à partir d'une ontologie de domaine est réalisée en utilisant les techniques de génération du langage naturel. Ce qui impose certaines restrictions comme les noms des propriétés qui doivent être sous forme de verbes. Les réponses proposées comportent une réponse correcte et plusieurs réponses fausses générées à partir des relations sémantiques entre les différents éléments de l'ontologie (la dimension syntaxique n'est pas prise en considération). La principale contribution de ce travail réside dans la proposition de stratégies qui servent à générer les réponses correctes et les réponses fausses du questionnaire. Ces stratégies concernent trois catégories : les classes, les propriétés et la hiérarchie de l'ontologie (sans les instances). Les questions sont générées à partir des connaissances affirmées et les connaissances inférées. Cette proposition a été optimisée et implémentée en tant que plugin de Protégé par (Tosic & Cubric, 2009).

L'objectif du travail de (Cubric & Tosic, 2011) est d'automatiser l'évaluation dans le contexte du e-learning mais cela peut se généraliser à l'évaluation des ontologies dans le cadre de l'ingénierie ontologique. Les auteurs proposent l'utilisation des annotations pour la génération des questions ainsi que l'interprétation sémantique du mapping entre l'ontologie de domaine et les questions. L'interprétation est basée sur les templates des questions basées sur la taxinomie de Bloom.

Dans (Abacha et al., 2013), l'approche proposée repose sur la génération de questions en langue naturelle à partir des éléments ontologiques à valider. Ces questions sont soumises à un expert. L'évaluation des réponses permet de décider de la validité des éléments ontologiques ou de leur modification dans le cas d'une évaluation négative. Cette approche exploite les techniques du traitement automatique de la langue (TAL) pour générer les questions et corriger l'ontologie selon les réponses des experts. Les questions générées sont soumises à une phase d'optimisation, utile surtout dans le cas d'ontologies de grande taille. Les réponses des experts sont divisées en deux catégories : une partie booléenne et une partie textuelle libre. Les éléments ontologiques invalidés par un expert sont supprimées de l'ontologie puisque son avis est considéré comme infaillible. S'il y a une adéquation parfaite, le nombre de questions générées est réduit à néant après plusieurs itérations entre le système et l'expert.

Le travail de (Alsubait et al., 2014) s'intéresse à évaluer le coût de la génération de questions à partir des ontologies à des fins éducationnelles. En effet, deux options s'offrent aux enseignants en termes de source de questions : les textes et les ontologies. Les auteurs soulignent que l'un des avantages des ontologies est qu'elles permettent de générer des questions sur des connaissances implicites grâce aux possibilités de raisonnement sur les ontologies. Toutefois, la disponibilité d'ontologies à valeur éducationnelle relatives à plusieurs domaines d'étude n'est pas garantie. Les auteurs proposent une théorie basée sur la similarité pour évaluer la difficulté des QCM. Le but est de contrôler la difficulté en variant la similarité entre la réponse correcte et les réponses fausses.

Le travail de (Richard et al., 2015) propose la méthode LOVMI (Les Ontologies Validées par Méthode Interactive) pour la validation d'ontologies, et en particulier l'ontologie ONTOPSYCHIA développée pour le module « facteurs sociaux et environnementaux ». Cette méthode propose de prendre en considération six dimensions dont la validation sémantique.

Les acteurs du domaine de l'ontologie visualisent l'arborescence conceptuelle de l'ontologie (y compris les axiomes et les relations). Des séances de validation se déroulent par groupes de deux pour discuter du contenu de l'ontologie. Il y a des échanges d'idées et des discussions sur les concepts de l'ontologie, les conversations sont enregistrées et les commentaires enregistrées sur WEBPROTÉGÉ (Protégé en ligne) sous forme d'annotations.

Parmi ces travaux, l'approche proposée par (Abacha et al., 2013) est celle qui présente un certain nombre de similarités avec notre proposition. La partie commune concerne la génération des questions sur la base des éléments ontologiques. Cependant, leur travail préconise l'utilisation des techniques du TAL, l'évaluation est réalisée par les experts du domaine et les réponses aux questions sont soient booléennes soit textuelles. Pour notre part, nous avons conçu des patrons de questions pour chaque élément ontologique, nous avons prévu deux évaluations, une par les concepteurs des ontologies et une autre par les experts du domaine. Les réponses aux questions varient selon les éléments ontologiques en fonction des critères de qualité cités dans la section 3.4.

5 Conclusion

La validation des ontologies par les experts du domaine est cruciale dans tout processus de développement d'une ontologie. Ce travail s'inscrit dans la continuité d'un travail antérieur, où nous avons proposé une approche supportée par un outil de génération automatique d'un questionnaire à partir d'une ontologie à évaluer. Et cela, dans le but de faciliter le travail du concepteur de l'ontologie d'une part et d'autre part, offrir aux experts du domaine la possibilité d'évaluer une ontologie sans avoir à maîtriser les langages de représentation des connaissances.

Nous envisageons de poursuivre ce travail en prenant en considération les aspects suivants:

- Tester l'outil de génération automatique des questionnaires à grande échelle auprès des concepteurs d'ontologies dans un premier temps puis auprès des experts du domaine dans un deuxième temps.
- Prendre en charge les nouveautés de la version 2 du langage OWL (Cardinalité qualifiée, asymétrie de propriétés, réflexivité de propriétés, restrictions sur les types de données, annotations d'axiomes, nouveaux profils, etc.)
- Prendre en charge plusieurs langues naturelles pour la génération de questionnaires en langue naturelle autre que l'Anglais.
- Explorer les possibilités offertes par les techniques de génération automatique de texte en langue naturelle (NLG) pour la dérivation de questionnaires à partir des ontologies

Références

ABACHA A. B. DA SILVEIRA M. & PRUSKI, C. (2013). Une approche pour la validation du contenu d'une ontologie par un système à base de questions/réponses. In IC-24èmes Journées francophones d'Ingénierie des Connaissances , Lille, France.

ALSUBAIT T. PARSIA B. & SATTLER U. (2014). Generating Multiple Choice Questions From Ontologies: Lessons Learnt. In OWLED, p. 73-84.

BRANK J. GROBELNIK M. & MLADENIC D. (2005). A survey of ontology evaluation techniques. Proceedings of Data Mining and Data Warehouses (SiKDD), Ljubljana, Slovenia.

BURTON-JONES A. STOREY V. SUGUMARAN V. (2005). A semiotic metrics suite for assessing the quality of ontologies, Data and Knowledge Engineering 55(1), p. 84–102.

CUBRIC M. & TOSIC M. (2011). Towards automatic generation of e-assessment using semantic web technologies. International Journal of e-Assessment, 1(1), p. 1-9.

DALKEY N. C. & HELMER O. (1963). An experimental application of the Delphi method to the use of experts, Management Science, 9(3), p 458-467.

ESCHENBACH C. & GRUNINGER M. (2008). Formal Ontology in Information Systems. Proceeding of Fifth Conference, FOIS, IOS Press.

FERNANDEZ-LOPEZ M. GOMEZ-PEREZ A. JURISTO N. (1997). METHONTOLOGY: From Ontological Art, Towards Ontological Engineering, AAAI Symposium on Ontological Engineering, Stanford, USA.

FOX M. S. BURBUCEANU M. & GRUNINGER M. (1996). An organization ontology for enterprise modelling: preliminary concepts for linking structure and behavior, Computers in Industry 29, p123–134.

GANGEMI A. CATENACCI C. CIARAMITA M. & LEHMANN J. (2005). Ontology evaluation and validation: An integrated formal model for the quality diagnostic task, Technical report, Laboratory of Applied Ontologies – CNR, Rome, Italy.

GOMEZ-PEREZ A. (2004 a). Ontology Evaluation, Handbook on Ontologies, p. 251-274.

GOMEZ-PEREZ A. (2004 b). Ontology Evaluation. Thèse de Doctorat, Faculté d'Informatique Université de Madrid, 2004.

GOMEZ-PEREZ A. (2004c). Ontology evaluation. In Steffen Staab and Rudi Studer, editors, Handbook on Ontologies in Information Systems, International Handbooks on Information Systems, chapter 13, p. 251–274.

GRUBER T. R. (1993). A translation approach to portable ontology specifications. Knowledge acquisition. 5(2), p. 199-220.

Gruber T. R. (1995). Towards principles for the design of ontologies used for knowledge sharing,

International Journal of Human-Computer Studies, 43(5/6), p. 907–928.

HOLSAPPLE C. & JOSHI K.D. (2005). A collaborative approach to ontology design. Communications of ACM, 45(2), p. 42–47.

HSU C. C. & SANDFORD B. A. (2007). The Delphi Technique:Making Sense Of Consensus, Practical Assessment, Research & Evaluation, 12, p.1-8.

LOVRENČIĆ S. & ČUBRILO M. (2008). Ontology Evaluation –Comprising Verification and Validation. Proceedings of CECIIS, the 19th Central European Conference on Information and Intelligent Systems, FOI, Varaždin, Croatia, p. 657-663.

OBRST L. CEUSTERS W. MANI I. RAY S. & SMITH B. (2007). The evaluation of ontologies, In Christopher J.O. Baker and Kei-Hoi Cheung, editors, Revolutionizing Knowledge Discovery in the Life Sciences, chapter 7, Springer, p.139–158.

PAPASALOUROS A. KANARIS K. & KOTIS K. (2008). Automatic Generation Of Multiple Choice Questions From Domain Ontologies. In e-Learning, p. 427-434.

RICHARD M. AIME X. KREBS M. O. & CHARLET J. (2015). LOVMI: vers une méthode interactive pour la validation d'ontologies. In 26es journées francophones d'Ingénierie des Connaissances (IC), Rennes, France.

SHETH A. TARTIR S. & BUDAK A. (2010). Ontological evaluation and validation. PhD thesis, Wright State University Main Campus, USA.

SUAREZ –FIGUEROA M. GOMEZ-PEREZ A. & FERNANDEZ-LOPEZ M. (2012). The NeOn Methodology for Ontology Engineering, Book Chapter in Ontology Engineering in a Networked World, Springer, Berlin Heidelberg, p. 9-34.

SURE Y. & STUDER R. (2002). On-To-Knowledge methodology. In Davies, J. et al. eds. On-To-Knowledge: Semantic Web enabled Knowledge Management. J. Wiley and Sons.

TOSIC M. & CUBRIC M. (2009). SeMCQ–Protégé Plugin for Automatic Ontology-Driven Multiple Choice Question Tests Generation. In Proceedings of the 11th International Protege Conference. Stanford Center for Biomedical Informatics Research, USA.

ZEMMOUCHI-GHOMARI L. & GHOMARI A.R. (2009). Reference Ontology. International IEEE Conference on Signal-Image Technologies and Internet-Based System, Marrakech, Morocco, p. 485-491.

ZEMMOUCHI-GHOMARI L. & GHOMARI A.R. (2014). Chapter sixteen a new approach for human assessment of ontologies. Business Intelligence and Mobile Technology Research: An Information Systems Engineering Perspective, p. 226-233.

Mesurer la qualité des systèmes de catégories de blogs

Ivan Garrido-Marquez, Jorge Garcia Flores,
François Lévy, and Adeline Nazarenko

LIPN, CNRS & Université Paris 13 – Sorbonne Paris Cité, 99 Jean-Baptiste Clément, 93430 Villetaneuse, France
{garridomarquez, jgflores, fl, adeline.nazarenko}@lipn.univ-paris13.fr

Résumé : Dans le monde prolifique de la blogosphère, retrouver une information pertinente est un défi. Pour aider leurs lecteurs, les auteurs de blog annotent souvent leurs billets avec des catégories mais ce système de catégories est difficile à maintenir quand le blog évolue au fil du temps. Pour guider la révision des annotations des blogs, nous proposons ici deux métriques, l'équilibre et le coût d'accès. Elles permettent de mesurer la qualité formelle des systèmes de catégories considérés en tant que systèmes d'index. Nos expériences montrent qu'en raisonnant sur les caractéristiques du système d'annotation et sur les mesures d'équilibre et de coût qu'on obtient par calcul, on a une vue synthétique sur le système de catégories à un instant t et que cela donne des indications sur les améliorations qu'on peut y apporter. Ces mesures sont conçues pour être intégrées dans un outil de diagnostic et de révision interactive des annotations de blogs que nous développons.

Mots-clés : Mesure de qualité, Catégorisation, Équilibre, Entropie, Blogs, Annotation

1 Introduction

Les blog sont des sites web utilisés pour la publication d'articles d'actualité présentant des informations ou les réflexions d'un ou plusieurs auteurs sur un sujet donné. Ces « billets », qui sont datés et signés, sont présentés par ordre antéchronologique. On estime que la barre des 150 millions de blogs a été franchie en 2010 (Chapman, 2011) et c'est devenu un moyen de communication et d'influence essentiel pour les individus, les entreprises et les médias.

Dans un monde aussi prolifique, retrouver une information pertinente est un défi. Pour aider leurs lecteurs, les auteurs de blogs cherchent généralement à annoter leurs billets en leur associant des catégories et/ou des « tags » (mots-clefs). L'ensemble des annotations aide le lecteur à retouver l'information pertinente en lui permettant de *naviguer* dans le blog (dans le plan de classement ou par proximité thématique) ou de *formuler des requêtes* et de retrouver directement tous les billets associés à un thème donné.

Avec le temps, le nombre de billets augmente et les thèmes évoluent au gré de l'actualité, si bien que le système de catégories proposé à un instant donné peut être beaucoup moins adéquat quelques mois ou quelques années plus tard. Nous nous intéressons ici au rôle d'index que jouent les systèmes de catégories de blogs : est-ce que le jeu de catégories proposé permet à un lecteur de retrouver efficacement les billets qui l'intéressent dans le blog ?

Dans cette optique, nous proposons des métriques permettant de mesurer la qualité d'un système de catégories de blogs. Il ne s'agit pas de vérifier l'adéquation des annotations au contenu des billets (nous supposons que les annotations, souvent posées par les auteurs des billets, sont localement pertinentes) mais de mesurer la qualité globale du système de catégories

* Ce travail s'inscrit dans le cadre des travaux sur l'accès aux contenus développé dans l'axe « Analyse sémantique computationnelle » du Labex EFL (ANR-10-LABX-0083).

comme système d'indexation. On observe en effet que les auteurs, qui annotent leurs billets un à un, n'ont guère de vision globale sur l'efficacité du système de catégories qu'ils construisent de manière de manière incrémentale. Les métriques permettent de détecter quand le système de catégories perd en efficacité et de suggérer les modifications à y apporter.

Après une revue des travaux analysant la qualité des systèmes d'annotation pour l'accès à l'information (sec. 2), nous présentons FLOG, le corpus de blogs que nous avons constitué (sec. 3). Nous introduisons deux mesures complémentaires pour apprécier l'équilibre du système d'annotation et le coût d'accès à l'information pour le lecteur (sec. 4) et nous présentons les résultats obtenus pour les blogs de FLOG (sec. 5). La discussion (sec. 6) montre pour finir comment nous prévoyons d'utiliser ces mesures pour faire le diagnostic des systèmes de catégories des blogs et proposer des mesures correctives aux annotateurs.

2 Etat de l'art : la qualité des systèmes d'annotation

Dans la littérature, la qualité des systèmes d'annotation est d'abord étudiée sous l'angle de l'adéquation entre l'annotation et la sémantique du texte annoté. Des études de ce type sont réalisées depuis longtemps sur l'indexation de ressources bibliographiques, notamment sur le domaine médical : Funk & Reid (1983) ont testé la cohérence d'indexation de 9 catégories définies à partir des en-têtes, sous-titres et concepts de MESH dans un ensemble d'articles en vue d'améliorer la fiabilité des stratégies de recherche. Leininger (2000) analyse en détail la cohérence inter-indexeur dans la base de données PsycINFO en utilisant deux mesures proposées par Hooper (1965) et Rolling (1981). Wilczynski & Haynes (2009) s'intéressent également à la capacité discriminante du vocabulaire d'indexation pour mesurer la qualité d'un système d'indexation ou d'annotation, tandis que (Cohen, 1960; Mathet *et al.*, 2012) ont évalué la qualité des systèmes d'indexation basé sur un vocabulaire contrôlé (Funk & Reid, 1983; Leininger, 2000; Wilczynski & Haynes, 2009). Nous ne nous intéressons pas directement à la qualité des annotations, que nous supposons bonne [1]. Nous cherchons à mesurer la qualité d'un système d'annotation considéré comme un outil d'accès à l'information. Dans cette perspective, il faut considérer le système d'annotation comme l'association d'un jeu de catégories, d'une collection de documents et de l'ensemble des liens d'annotations qui relient les catégories aux documents.

3 FLOG, un corpus de blogs français

Ce corpus (Garrido-Marquez *et al.*, 2016) contient 20 blogs différents, 25 000 billets et 11 millions des mots. Les blogs relèvent de 4 grands thèmes (cuisine, jeux video, technologie et droit) et le corpus couvre une période de 10 ans. Les billets sont annotés par leurs auteurs avec des catégories et/ou des tags.

On observe sur ce corpus que les habitudes d'annotation varient d'un blog à l'autre. Nous nous intéressons ici spécifiquement aux annotations de type catégories. Le nombre de catégories par blog varie entre 4 et 91 et pas nécessairement en proportion du nombre de billets, puisque le nombre moyen de billets par catégorie va de 2 à 64 (à l'arrondi près).

1. Nous faisons l'hypothèse que les auteurs de billets savent dans quelle(s) catégorie(s) les ranger ou disposent d'outils pour les aider à le faire, à partir de l'analyse du contenu du billet.

Les annotateurs utilisent les systèmes de catégories de différentes manières. Dans les *systèmes mono-catégoriels*, les catégories sont utilisées de manière exclusive et un billet n'est annoté que par une seule catégorie. Il y a 6 blogs de ce type dans FLOG. Dans les *systèmes multi-catégoriels*, un même billet peut être associé à plusieurs catégories. Enfin, les systèmes de catégories peuvent être structurés hiérarchiquement (*systèmes hiérarchiques*). Le corpus FLOG ne contient aucun blog de cet type, même si `technologie2` s'en rapproche [2].

4 Mesurer l'équilibre et le coût d'accès d'un système d'annotation

L'*équilibre* d'un système de catégories caractérise l'information intrinsèquement contenue dans ce système, vue comme une distribution de probabilité sur les catégories.

S'agissant d'un blog, la distribution ne comporte qu'un ensemble fini $\mathcal{F}$ d'événements qui sont des unions d'événements élémentaires. Dans l'analyse de Shannon (1948), la quantité d'information $I(e)$ recelée par un événement particulier e est $-\log_b(P(e))$ et l'entropie est la valeur espérée de cette quantité d'information. En notant $\mathcal{A}$ les événements élémentaires de $\mathcal{F}$, on peut mesurer l'entropie H :

$$H = E_{\mathcal{F}}[I(e)] = \sum_{e \in \mathcal{A}} P(e)I(e) = -\sum_{e \in \mathcal{A}} P(e)\log_b(P(e)) \tag{1}$$

Pour les systèmes mono-catégoriels, l'adaptation est directe : un événement élémentaire est une catégorie et l'on utilise la fréquence de la catégorie comme sa probabilité. La mesure d'entropie ne suffit cependant pas pour comparer deux blogs ou deux versions différentes d'un système de catégories, parce que la valeur maximale dépend du nombre de catégories (le meilleur système aurait une catégorie par billet !). L'*équilibre* (Pielou, 1966) rapporte donc l'entropie calculée par l'équation 1 à l'entropie maximum susceptible d'être obtenue avec le même nombre de catégories, soit pour un blog x comportant N billets et n catégories x_i ($i = 1 \ldots n$) :

$$H(x) = -\sum_{i=1}^{n} \frac{|x_i|}{N}\log_b(\frac{|x_i|}{N}) \qquad Equilibre(x) = \frac{H(x)}{\max(H(x))} = \frac{H(x)}{\log_b(n)} \tag{2}$$

Dès lors qu'un billet peut être annoté par plusieurs catégories, ce qui est le cas dans 2/3 des blogs du corpus FLOG, les catégories ne représentent plus des événements élémentaires car certains billets sont décrits par une *combinaison de catégories*. On peut montrer cependant que les événements élémentaires sont calculables à partir des combinaisons de catégories sans négation, ce qui permet d'utiliser la mesure d'équilibre ci-dessus.

Nous cherchons également à apprécier le *coût d'accès* aux billets au sein d'un blog indépendamment des interfaces de navigation proposées aux utilisateurs. Nous considérons pour cela le schéma de base où un lecteur formule une requête composée d'une ou plusieurs catégories et reçoit en retour un ensemble de billets qu'il doit parcourir pour trouver celui qui l'intéresse. Le coût d'accès aux billets du blog b s'exprime comme la somme des coûts de composition de la requête ($cout_{req}$) et de sélection d'un billet dans l'ensemble des billets retournés ($cout_{doc}$) :

$$Cout(b) = cout_{req}(b) + cout_{doc}(b) \tag{3}$$

2. Comme les billets annotés par les catégories feuilles de l'arbre sont parfois aussi annotés par des catégories plus génériques, nous le considérons comme un système multi-catégoriel.

Blog	**Type**	**T**$_{voc}$	**N**$_{doc}$	**Équilibre**	C_{cat}	C_{doc}	**Coût**
cuisine1	Mono	60	452	0,69	60	7,53	67,53
cuisine2	Mono	26	927	0,82	26	35,65	61,65
jeuxvideo1	Multi	43	1422	0,82	140,76	46,85	187,62
technologie1	Multi	56	1423	0,63	59,81	229,88	289,69
technologie5	Multi	16	132	0,73	33,59	34,32	67,92
droit1	Mono	4	243	0,72	4	60,75	64,75
jeuxvideo2	Multi	33	1234	0,86	129,93	6,17	136,11
technologie2	Multi	38	243	0,89	69,87	7,62	77,50
jeuxvideo3	Multi	91	5486	0,76	91	335,04	426,04
jeuxvideo4	Multi	40	1501	0,80	87,01	30,01	117,02
droit2	Multi	48	931	0,64	76,65	121,63	198,29
cuisine3	Mono	50	395	0,83	50	7,90	57,90
technologie3	Multi	41	343	0,84	53,28	24,08	77,37
droit3	Multi	13	283	0,73	17,76	52,66	70,42
cuisine4	Mono	25	1561	0,69	25,00	62,44	87,44
droit4	Multi	15	1572	0,57	43,13	161,21	204,34
technologie4	Mono	12	573	0,74	12,00	47,75	59,75
jeuxvideo5	Multi	37	1134	0,90	105,28	11,30	116,58
technologie6	Multi	16	374	0,85	95,70	16,20	111,91
jeuxvideo6	Multi	18	184	0,93	18,18	11,71	29,90

TABLE 1 – Caractéristiques des blogs du corpus FLOG et mesures d'équilibre et de coûts

Pour un blog ayant un système d'annotation multi-catégoriel (dont le mono-categoriel est pour ce calcul un cas particulier), il faut tenir compte de la taille T_{voc} du vocabulaire de catégories proposé. Formuler une requête r de longueur l revient à choisir successivement l catégories parmi les T_{voc} disponibles, ce qui a un coût $cout_{req}(r) = \sum_{i=0}^{l-1}(T_{voc}(b) - i)$. Le coût $cout_{req}(b)$ de composition des requêtes du blog est l'espérance de ce coût de composition pour une requête. De même, le coût de sélection du document est l'espérance du nombre $Eff_c(r)$ de documents ramenés par r.

$$C_{multi}(b) = E_{r \in Req}\Big(\sum_{i=0}^{l(r)-1} (T_{voc}(b) - i) + Eff_c(r)\Big) \tag{4}$$

Dans un système hiérarchique, les billets sont généralement décrits par une seule catégorie mais le choix de la catégorie est guidé par la structure arborescente. Considérons pour simplifier un arbre complet et équilibré de degré d. La hauteur h de cet arbre est $h = log_d(T_{voc})$. Pour sélectionner une catégorie feuille, il faut choisir h fois une catégorie parmi les d disponibles à chaque niveau. En notant Eff_c l'effectif moyen des catégories feuilles, on a :

$$C_{arbre}(b) = log_d(T_{voc}(b)).d + Eff_c(b) \tag{5}$$

5 Résultats expérimentaux

Le tableau 1 présente les mesures d'équilibre et de coûts pour chacun des 20 blogs du corpus FLOG. L'équilibre, qui varie entre $0,57$ et $0,93$, donne une idée rapide de la distribution des billets dans les catégories des différents blogs : on voit par exemple que les billets de `jeuxvideo6` sont plus uniformément distribués que ceux de `cuisine1` [3] ; on voit aussi que

3. De fait, `cuisine1` présente une catégorie majoritaire associée à près de 25% des billets ; près de 50% des billets se retrouvent dans les 3 principales catégories et 85% des billets se concentrent sur seulement 15 catégories.

les mesures de coûts sont extrêmement variables (de 20 à 420) et qu'on peut avoir un coût d'accès élevé eu égard au nombre de documents même pour un blog relativement équilibré (ex. `technologie2`), preuve que les deux mesures sont complémentaires.

L'évolution de ces mesures dans le temps est également intéressante. Elle montre que les auteurs, mêmes s'ils veillent à l'adéquation des catégories qu'ils posent au contenu des billets (analyse locale), ne se rendent pas toujours compte de la qualité de l'index créé par ces catégories (analyse globale). Dans la majorité des cas, on observe en effet que l'équilibre se dégrade au cours du temps. Les graphiques de la figure 1 le montrent pour les blogs `jeuxvideo3` et `cuisine4` (courbes noires). Dans certains cas, l'ajout de nouvelles catégories limite la dégradation de l'équilibre (ex. `jeuxvideo3`) mais dans d'autres, cela paraît au contraire contre-productif (ex. `cuisine4`)[4]. On fait la même constatation sur les graphes de coût.

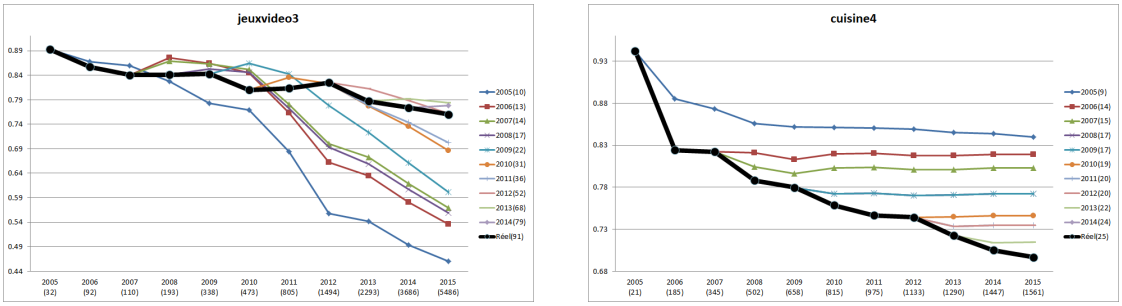


FIGURE 1 – Evolution de l'équilibre (courbes noires) et projections de l'équilibre à jeu de catégories constant (courbes de couleur)

6 Discussion : vers un diagnostic des systèmes de catégories de blogs

L'analyse longitudinale d'un corpus de blogs et les mesures que nous avons introduites pour mesurer la qualité de l'indexation proposée par les systèmes de catégories montrent que cette qualité d'indexation est difficile à contrôler par les auteurs de blogs qui se préoccupent prioritairement de la qualité des annotations qu'ils posent au regard des contenus qu'ils cherchent à annoter et qui n'ont qu'une vision locale des blogs qu'ils annotent.

Nous considérons qu'une plateforme de gestion de blog doit non seulement proposer des fonctionnalités d'annotation – permettre de poser des tags/catégories sur les billets, avec éventuellement des outils d'aide à l'annotation – mais qu'elle doit aussi offrir des outils de diagnostic permettant d'apprécier et d'améliorer la qualité d'un système de catégories en termes d'indexation. Les mesures proposées dans cet article permettent de fonder ce type de diagnostic.

Le suivi des mesures d'équilibre et de coût permet de *détecter* une dégradation de la qualité d'indexation qui rend les billets difficiles d'accès pour le lecteur et d'alerter l'auteur du blog. Il faut ensuite *localiser* les catégories défaillantes et proposer des mesures correctives à l'utilisateur. Les indications dépendent des mesures obtenues. Si l'équilibre est faible, on peut soit décomposer les grosses catégories en sous-catégories soit regrouper des petites catégories. Si le

4. Les courbes en couleur montrent les mesures d'équilibre qu'on aurait obtenues si on avait gardé un jeu de catégories inchangé : les courbes bleu clair retracent ainsi l'évolution de l'équilibre qu'on aurait obtenu en conservant le jeu de catégories de 2005 jusqu'en 2015.

coût d'accès aux documents est élevé, il faut globalement affiner la granularité des catégories, soit en décomposant les catégories existantes, soit en introduisant des catégories indépendantes (le système devient multi-catégoriel) pour réduire le coût d'accès aux documents sans augmenter trop le nombre de catégories. Quand le coût d'accès aux catégories est élevé, il faut une réorganisation globale du système de catégories en système multi-catégoriel ou hiérarchique. Il arrive aussi que certains systèmes de catégories soient inefficaces car redondants – c'est le cas du blog `technologie5` – et on s'en rend compte quand on observe qu'on a un système multi-catégoriel avec des coût d'accès aux catégories et aux documents tous les deux élevés.

Établir la liste des corrections à proposer à l'auteur nécessite cependant de compléter l'analyse globale qui est faite ici par une analyse plus détaillée. Il faut localiser les catégories les plus grosses, les plus petites ou les plus redondantes. Il faut également tenir compte de l'âge des catégories et de leur taux d'activité : il est inutile de regrouper des catégories qui sont en croissance forte mais peut-être urgent, à l'inverse, de décomposer une catégorie importante qui continue à se développer. Il faut enfin prioriser les corrections à faire et tenir compte du coût induit par la correction (combien de billets faut-il réannoter ?).

C'est l'auteur qui choisit ou pas de *réparer* le système de catégories à partir des propositions de correction qui lui sont faites mais on voit que les mesures proposées ici permettent d'établir un diagnostic et de l'éclairer sur la qualité de l'index que constitue son système de catégories.

Références

CHAPMAN C. (2011). A brief history of blogging. `http://www.webdesignerdepot.com/2011/03/a-brief-history-of-blogging/`. [Marketing, Web Design, WordPress · Mar 14, 2011].

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

FUNK M. E. & REID C. A. (1983). Indexing consistency in MEDLINE. *Bulletin of the Medical Library Association*, **71**(2), 176–183.

GARRIDO-MARQUEZ I., AUDIBERT L., GARCÍA-FLORES J., LÉVY F. & NAZARENKO A. (2016). A French Weblog Corpus for New Insights on Blog Post Tagging. In A. M. ORTIZ & C. PÉREZ-HERNÁNDEZ, Eds., *CILC2016. 8th International Conference on Corpus Linguistics*, volume 1 of *EPiC Series in Language and Linguistics*, p. 144–158 : EasyChair.

HOOPER R. S. (1965). *Indexer consistency tests : origin, measurement, results, and utilization*. Rapport interne, IBM Corporation, Bethestda, MD.

LEININGER K. (2000). Interindexer consistency in psycinfo. *Journal of Librarianship and Information Science*, **32**(1), 4–8.

MATHET Y., WIDLÖCHER A., FORT K., FRANÇOIS C., GALIBERT O., GROUIN C., KAHN J., ROSSET S. & ZWEIGENBAUM P. (2012). Manual Corpus Annotation : Giving Meaning to the Evaluation Metrics. In *International Conference on Computational Linguistics*, p. 809–818, Mumbaï, India.

PIELOU E. (1966). The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology*, **13**, 131 – 144.

ROLLING L. (1981). Indexing consistency, quality and efficiency. *Information Processing & Management*, **17**(2), 69–76.

SHANNON C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 379–423, 623–656.

WILCZYNSKI N. L. & HAYNES R. B. (2009). Consistency and accuracy of indexing systematic review articles and meta-analyses in medline. *Health Information & Libraries Journal*, **26**(3), 203–210.

Recommandation de ressources pédagogiques au sein d'un système de systèmes d'information

Mohamed Ali Ben Ameur[1], Majd Saleh[1], Marie-Hélène Abel[1] et Elsa Negre[2]

[1] Sorbonne universités, Université de technologie de Compiègne
CNRS, HEUDIASYC, UMR 7253**,** CS 60 319,
60203 Compiègne Cedex, France,
{mohamed-ali.ben-ameur, majd.salah, marie-helene.abel}@utc.fr
[2] Université Paris-Dauphine, PSL Research Universities,
LAMSADE UMR CNRS 7243,
75016 Paris, France,
elsa.negre@dauphine.fr

Résumé : Avec le développement des Technologies de l'Information et de la Communication, les organisations sont confrontées à une grande quantité d'informations issues de nombreux systèmes. Identifier une ressource d'information pertinente en fonction d'un contexte précis devient un vrai challenge. Dans le cadre de notre travail nous nous intéressons aux écosystèmes apprenants et à la recommandation de ressources pédagogiques. Nous avons fait le choix de modéliser un écosystème apprenant comme un système de systèmes d'information (SoIS) au sein duquel nous introduisons un système de recommandations de ressources basé sur le vote des utilisateurs (apprenants, enseignants) de l'écosystème. Dans notre approche, nous tenons compte de qui recommande quoi à qui sur quel sujet, quand, comment et pourquoi. Nous le faisons au moyen du modèle de collaboration memorae-core2 mis en œuvre pour la conception du SoIS support à l'écosystème apprenant. Dans cet article, nous précisons notre problématique, justifions notre approche SoIS et présentons la mise en œuvre du modèle memorae-core2 pour la conception du SoIS-memorae.
Mots-clés : système de systèmes d'information, écosystèmes apprenants, système de vote, système de recommandations.

1 Introduction

Aujourd'hui avec l'arrivée des Technologies de l'Information et de la Communication (TIC), les apprenants évoluent dans un écosystème apprenant qui peut être défini comme un modèle écologique d'apprentissage et d'enseignement en ligne (Frielick, 2004). Il peut être vu comme un espace d'apprentissage virtuel dans lequel les technologies qui concourent à l'apprentissage sont utilisées, dans le but de favoriser les interactions entre communautés d'utilisateurs et de contenu. Une communauté est constituée de personnes en interaction qui partagent, utilisent des informations, des connaissances sur des centres d'intérêts communs. Dans cette introduction, nous exposons le contexte social et scientifique de notre recherche.

1.1 Contexte social

Dans le contexte de « l'apprentissage ensemble », de nombreux systèmes d'information (SI) sont utilisés par les apprenants. Ces systèmes fournissent des ressources hétérogènes (vidéo, texte, e-book, forum en ligne, etc.) aux différents utilisateurs (étudiants, enseignants). Selon Guy et Carmel (2011) la multitude de ressources, de relations et d'interactions peut conduire les utilisateurs à subir une surcharge informationnelle qui les rend incapables d'assimiler les informations disponibles. Afin de réduire cette surcharge, il serait utile d'offrir une aide aux utilisateurs afin qu'ils puissent choisir les ressources susceptibles d'être les plus pertinentes dans une situation donnée. Une des voies possibles allant dans ce sens concerne le partage d'information et les systèmes de vote bien connu des internautes. Il devient donc intéressant

de s'interroger sur l'association de ces différents moyens dans le cadre d'un écosystème apprenant afin de produire des recommandations ciblées.

1.2 Contexte scientifique

La recherche d'information dans le contexte des Environnements Informatiques pour l'Apprentissage Humain (EIAH) demeure un défi à relever. Dans la littérature, ce défi est généralement abordé en considérant le profil des apprenants, notamment dans les travaux sur le filtrage d'information afin de proposer aux apprenants des documents pertinents, en particulier : l'approche par contenu (Pazzani et Billsus, 2007), l'approche à base de connaissances (Burke, 1996) et l'approche par filtrage collaboratif (Goldberg, 1992). Dans le cadre de cette dernière approche, au filtrage collaboratif peut être associée la prise en compte du profil des utilisateurs apprenants (qui consulte quoi ?). Bien que plus précis, ces systèmes ne tiennent pas compte du contexte de collaboration et de la possibilité de partager des ressources issues de différents systèmes. Dans le contexte de recommandations au sein d'un écosystème apprenant (EA), nous pensons qu'il est nécessaire de prendre en considération :

- la volonté de partager des ressources avec une communauté afin d'atteindre un objectif commun,
- le fait que les ressources partagées peuvent provenir de différents SI.

Afin de répondre à ces nécessités, nous proposons de considérer un écosystème apprenant comme un système de systèmes d'information, SoIS, développé à partir d'un modèle de collaboration et comprenant un système de recommandations basé sur le vote des utilisateurs.

2 Écosystème apprenant versus SoIS

Considérer un écosystème apprenant comme un SoIS vise à simplifier la gestion des ressources pédagogiques provenant de différents systèmes d'information, et le contrôle du processus de partage des informations entre les apprenants. L'objectif est de minimiser le temps nécessaire pour capitaliser les ressources issues de différents systèmes d'informations.

2.1 Écosystème apprenant

A l'ère des technologies 2.0, le concept d'individu-plus (Perkins, 1995) prend tout son essor. L'apprenant n'évolue pas seul, individu solo, mais dans un écosystème apprenant comprenant l'apprenant lui-même, mais également son environnement physique et social : ses outils (bloc-notes, tablette, etc.), ses ressources (procédures, documentation, etc.), ses partenaires qui disposent eux-aussi d'une partie de la connaissance (enseignants, collègues, etc.).

Les écosystèmes numériques ont pour objectif de garantir le partage des connaissances au sein des organisations aussi rapidement et efficacement que possible (Price et Turnbull, 2007). Ils peuvent être considérés comme des plateformes support à la coopération, au partage et à l'accès aux connaissances afin de faciliter l'apprentissage (Serge et Nicole, 2016). Sous cet angle, ils peuvent donc servir de support à un écosystème apprenant, dans ce cas, nous les appellerons des écosystèmes apprenants numériques (EAN).

2.2 SoIS

Un système de systèmes (SoS) est une collection de systèmes dédiés qui regroupent leurs ressources et leurs capacités pour créer un nouveau système plus complexe qui offre plus de fonctionnalités et de performance que simplement la somme des systèmes constitutifs (Popper et Bankes, 2004). Différentes approches ont été proposées dans la littérature concernant la coordination des différents systèmes d'un SoS. Il existe principalement trois approches (Lozano, 2010) : *Leader / Follower*, *Virtual Structure* et *Behavioral Control*. Dans l'approche

Leader / Follower, un système *leader* permet aux systèmes composants de coopérer, de mener une tâche en collaboration (Dong, 2007).

Dans notre contexte, nous nous intéressons plus particulièrement à une catégorie de SoS : les SoIS. Carlsson et Stankiewicz (1991) les définissent comme des réseaux d'agents qui interagissent dans un domaine technologique spécifique afin de créer, diffuser et utiliser des technologies axées sur le savoir, l'information et le flux de compétences. Ainsi, un SoIS peut être considéré comme un macro-système d'information donnant accès aux informations distribuées dans les systèmes composants et offrant des fonctionnalités utilisant les informations accédées.

Les EAN et les SoIS partagent finalement un certain nombre de caractéristiques comme :

- la distribution des ressources de connaissance au sein de différents systèmes,
- la prise en compte de ressources hétérogènes de connaissances,
- le support au partage et à l'accès des ressources de connaissance et à la collaboration.

3 SoIS MEMORAe

Le SoIS MEMORAe a été développé comme support numérique à un EA. Il suit l'approche *Leader/Follower*. Le système *leader* a pour fonction d'orchestrer le SoIS. Il peut être vu comme une base de connaissances en lien avec les systèmes du SoIS permettant l'organisation, le partage, l'accès aux ressources des différents systèmes composants. Il a pour vocation de servir de support aux utilisateurs de l'EAN pour faciliter la collaboration et la prise de décision concernant l'objet de la collaboration.

3.1 Choix du modèle de collaboration memorae-core2

Le modèle de collaboration memorae-core2 a été développé dans le cadre du projet MEMORAe (Abel, 2015). Ce projet vise à gérer des ressources d'information hétérogènes au sein des organisations et à faciliter l'apprentissage organisationnel. La collaboration est considérée du point de vue du partage et de l'échange de ressources hétérogènes de connaissances entre collaborateurs utilisateurs autour du référentiel partagé.

Memorae-core2 modélise les utilisateurs en considérant une organisation comme un ensemble de membres qui interagissent. Chaque membre peut être un support à l'apprentissage organisationnel. A chaque utilisateur et groupe d'utilisateurs est associé un espace de partage ou seront rendues visibles/accessibles les ressources.

Les ressources sont modélisées comme des «vecteurs d'information». Chaque ressource est indexée par une où plusieurs clés d'indexation. Une clé d'indexation permet de rendre visible la ressource selon un concept du référentiel partagé dans un espace de partage.

Notons que dans memorae-core2, un vote est modélisé comme une ressource. Cette dernière permet à un utilisateur de préciser la pertinence d'une cible (ressource) pour un sujet défini dans le référentiel partagé au sein d'un espace de partage.

3.2 Architecture du SoIS MEMORAe

L'architecture du SoIS MEMORAe (Salah 2016) est basée sur le regroupement des informations et des ressources hétérogènes de différents systèmes d'information. Ces systèmes sont autonomes et fonctionnent séparément les uns des autres. Chacun d'eux possède ses propres services / fonctions et bases de données.

Le SoIS est représenté comme un groupe de connecteurs de systèmes et de bases de données comprenant les informations (identifiant/mot de passe) liées aux connections aux systèmes composants.

Développé à partir de memorae-core2 le système *leader* du SoIS MEMORAe propose aux utilisateurs apprenants les fonctionnalités suivantes :

- accéder aux ressources provenant de différents systèmes d'information dans un emplacement centralisé,

- organiser les ressources autour d'un référentiel partagé présenté sous la forme d'une carte (ontologie d'application),
- partager les ressources dans différents espaces de partage,
- annoter/voter les ressources afin de mettre en évidence certaines idées liées aux ressources.

4 Recommandations au sein du SoIS MEMORAe

Au sein du SoIS MEMORAe, il est possible de partager des ressources pédagogiques issues de différents systèmes d'information. Ces ressources peuvent présenter pour un même sujet un degré de pertinence variable selon le niveau des apprenants. Ce degré de pertinence est établi à partir du vote des utilisateurs (apprenants, enseignants) de l'écosystème et est exploité au sein d'un système de recommandations. C'est le système *leader* qui met en œuvre les fonctionnalités de collaboration et de recommandation en s'appuyant sur le modèle memorae-core2.

L'approche suivie considère plusieurs aspects : modélisation de l'apprenant, modélisation du contenu d'une formation, modélisation des ressources, recommandation des ressources pédagogiques.

4.1 Modélisation du contenu de formation

Notre écosystème apprenant a pour objectif d'aider l'apprenant utilisateur à appréhender les concepts d'une formation et à faciliter les échanges et le transfert de connaissances au sein de l'écosystème autour d'un référentiel partagé. Ainsi, le contenu de cours est présenté sous la forme d'une cartographie de connaissance (ontologie d'application). Cette ontologie est une spécification de l'ensemble des notions utiles à une formation particulière.

4.2 Modélisation de l'apprenant

Le modèle de l'apprenant est un élément essentiel dans le système de recommandations. (Brusilovsky ,1996). Il permet de caractériser chaque apprenant dans le système à travers un profil. Nous souhaitons exploiter le profil de l'apprenant afin de lui offrir des mécanismes qui lui permettent de mieux gérer ses ressources pour une formation qu'il souhaite valider. Les caractéristiques du modèle retenu concernent l'identité (nom, prénom, etc.) (Cheung et al, 2003), les préférences (langue, style, etc.) (Cheung et al, 2003), l'état de connaissances de l'apprenant sur les notions liées à la formation (Piombo, 2007). Ces caractéristiques peuvent être renseignées de manière explicite (formulaire rempli par l'apprenant) ou implicite (établi à partir de l'analyse des activités de l'apprenant lors de la formation). Nous nous basons sur les travaux de (Mediani et al, 2015) pour le renseignement implicite.

Au final, notre modèle de l'apprenant pour une formation suivie peut être exprimé par :

$$M_{Apprenant} = \{M_{ID}, M_{Préf,} M_{Co}\}$$

Où M_{ID} représente l'ensemble des informations sur l'apprenant (nom, …), $M_{préf}$ est l'ensemble des préférences de l'apprenant (quel format ?,...) et M_{Co} est l'ensemble des connaissances préalables de l'apprenant en rapport avec la formation.

4.3 Modélisation des ressources

Dans notre modèle memorae-core2, une ressource pédagogique est une ressource particulière. Le modèle des ressources pédagogiques, $M_{ressourcePedagogique} = \{M_{ID}, M_{ConnaissancesPertinenceDifficulté,} M_{Objectifs}\}$, est composé de trois parties. La composante M_{ID} est décrite par les éléments suivants : le titre, la description, le type, l'auteur, la langue et le format. La composante $M_{ConnaissancesPertinenceDifficulté}$ résulte de l'indexation de ressource par les concepts de la formation qu'elle aborde. Chaque ressource a un degré de pertinence et une

difficulté associés à la connaissance qu'elle traite. Le degré de pertinence est calculé à partir du vote des utilisateurs de l'écosystème membre du groupe ayant accès à l'espace de partage où la ressource est visible. La difficulté précise un niveau d'accessibilité pour un public donné (faible, moyen, haut) et est défini par les auteurs de la ressource. La troisième composante $M_{Objectifs}$ permet de préciser le but principal de l'utilisation d'une ressource dans la communauté qui peut être théorique (lecture, consultation) ou pratique (exercice, problème, etc.).

Afin de permettre aux utilisateurs apprenants d'effectuer un vote participatif sur l'intérêt d'une ressource pédagogique sur une notion à appréhender nous nous appuyons sur le modèle memorae-core2 dans lequel le concept de vote est défini comme une ressource ayant une cible.

La cible est une *clé d'index* représentée par un concept faisant le lien entre une ressource, un concept qui indexe cette ressource et un espace de partage où cette ressource est visible/partagée. De la sorte, il devient possible de voter sur l'intérêt d'une ressource au sein d'une communauté pour un concept particulier. Le degré de pertinence d'une ressource pour une communauté sur un sujet/connaissance sera calculé par la moyenne pondérée des votes émis.

4.4 Module de recommandations de ressources pédagogiques

Notre module de recommandations de ressources pédagogiques se base sur les liens sémantiques existant entre les concepts de la formation, les ressources indexées en tenant compte du degré de pertinence calculé, le niveau de difficulté qu'elles présentent et le modèle de l'apprenant (identité, préférences, connaissances). L'objectif d'un apprenant en suivant une formation est d'acquérir les connaissances liées à cette dernière. Ainsi pour un apprenant *a* désirant appréhender une connaissance *c* (concept) ayant accès à l'espace de partage *s*, nous lui recommanderons des ressources du système qui sont :

- indexées par le concept *c* et/ou les concepts *sc* spécialisant *c*,
- accessible dans l'espace de partage de la communauté à laquelle appartient *a*,
- ayant un degré de pertinence supérieur à un seuil (elle a été jugée pertinente par les membres de la communauté),
- ayant une difficulté adaptée au niveau de connaissance de l'apprenant *a* pour le concept *c* qui indexe cette ressource.
- ayant des caractéristiques (format, style, etc.) similaires aux préférences de l'apprenant *a* pour le concept *c*.

La recommandation consiste à suggérer à l'apprenant des ressources issues des systèmes composants du SoIS jugées pertinentes sur le concept ***c*** par les membres de sa communauté et qui s'adaptent au mieux à ses préférences, et à son niveau de connaissance.

5 Conclusion

Dans le cadre de notre travail, nous nous intéressons aux écosystèmes apprenants et à la recommandation de ressources pédagogiques. Nous avons fait le choix de modéliser un écosystème apprenant comme un système de systèmes d'information (SoIS) au sein duquel nous introduisons un système de recommandations de ressources basé sur le vote des utilisateurs (apprenants, enseignants) de l'écosystème. Dans notre approche, nous tenons compte de la volonté de collaborer. Par conséquent, nous exploitons le modèle memorae-core2 pour répondre à la demande de questions telles que, qui collabore avec qui, comment, quand, pourquoi, sur quoi et où, etc.

Un premier prototype a été développé et est en cours de test auprès des étudiants postbac de l'Université de Technologie de Compiègne suivant un cours d'informatique.

6 Remerciements

Ce travail a été réalisé dans le cadre d'un stage de Master financé par le Labex MS2T (ANR-11-IDEX-0004-02) et s'appuie sur la plateforme développée dans le cadre du projet ECOPACK (ANR-13-ASTR-0026).

Références

ABEL M. H. (2015). Knowledge Map-Based Web Platform to Facilitate Organizational Learning Return of Experiences. Computers in Human Behavior, p. 960-966.

BRUN A., HAMAD A., BUFFET O. & BOYER A. (2010). Vers l'utilisation de relations de préférence pour le filtrage collaboratif. 17eme congrès francophone Reconnaissance des Formes et Intelligence Artificielle, RFIA 2010, Jan 2010, Caen, France.

BRUSILOVSKY P. (1996). Methods and techniques of adaptive hypermedia. User Model. User-Adapt. Intract. 6(2-3), p. 87-129.

BURKE R., HAMMOND K. & YOUNG B. (1996). Knowledge-based navigation of complex information spaces. In Proc of the 13th National Conference on Artificial Intelligence. Canada. p. 462-468.

CARLSSON B. & STANKIEWICZ R. (1991). On the nature, function and composition of technological systems. Journal of evolutionary economics, vol. 1, no. 2, p. 93-118.

CHEUNG B., HUI L., ZHANG J. & YIU S.M. (2003). SmartTutor: An intelligent tutoring system in web-based adult education, The Journal of Systems and Software, Vol. 68, No. 1, p. 11-25.

DONG H. & HUSSAIN F. K. (2007). Digital ecosystem ontology, in EmergingTechnologies and Factory Automation, 2007. ETFA. IEEE Conference on. IEEE. p. 814-817.

FRIELICK S. (2004). Beyond constructivism: An ecological approach to e-learning.

GOLDBERG D., NICHOLS D., OKI B. M. & TERRY D. (1992). Using collaborative filtering to weave an information tapestry. Communications of the ACM. p. 61– 70.

GUY I. & CARMEL D. (2011). Social recommender systems. In Proceedings of the 20th international conference companion on World Wide Web. ACM. p. 283-284.

JAMSHIDI M. (2008). System of Systems Engineering. Innovations for the 21st Century, Wiley Ser. in Systems Engin. John Wiley & Sons.

LOZANO R., SPONG M. W., GUERRERO J.A. & CHOPRA N. (2010). Controllability and Observability of Leader-Based Multi-agent Systems. 20 Jan 2010.

MEDIANI C., ABEL M. H. & DJOUDI M. (2015). Towards a Recommendation System for the Learner from a Semantic Model of Knowledge in a Collaborative Environment. 5th IFIP TC 5 International Conference, CIIA 2015, Saida, Algeria, May 20-21, 2015, Proceedings. IFIP Advances in Information and Communication Technology 456, p. 315-327.

NACHIRA F., DINI P., NICOLAI A., LE LOUARN M. & RIVERA LEON L. (2007). Digital Business Ecosystems. Luxembourg, Office of Official Publications of European Communities.

PAZZANI M. & BILLSUS D. (2007). The Adaptive Web, chapter Content-Based Recommendation Systems. Springer Berlin / Heidelberg. p. 325-341.

PERKINS D. N. (1995). L'individu-plus. Une vision distribuée de la pensée et de l'apprentissage. Revue française de pédagogie Année 1995 Volume 111 Numéro1. p. 57-71.

STRAFFIN P. D. (1980). Topics in the theory of voting. The UMAP expository monograph series. Birkhauser, Boston.

PIOMBO C. (2007). Modélisation probabiliste du style d'apprentissage et application à l'adaptation de contenus pédagogiques indexés par une ontologie. Thèse de doctorant.

POPPER S., BANKES S., CALLAWAY R. & DELAURENTIS D. (2004). System-of-Systems Symposium: Report on a Summer Conversation. Potomac Institute for Policy Studies, Arlington.

PRICE C. & TURNBULL D. (2007). The organizational challenges of global trends. A McKinsey global survey. McKinsey Quarterly.

SALEH M., ABEL M. H., MISSERI V., MOULIN C. & VERSAILLES D. (2016). Moving Integration of Brainstorming Platform in a System of Information Systems. MEDES'16, Biarritz, France.

SERGE A. & NICOLE K. (2016). Les écosystèmes numériques Intelligence collective, développement durable, inter-culturalité, transfert de connaissance.

Du langage naturel à la connaissance il n'y a qu'un pas : SWIP

Mathilde Lannes[1], Fabien Amarger[1], Nicolas Seydoux[1,2], Nathalie Hernandez[1]

[1] IRIT UMR 5505, UT2J Maison de la Recherche, 5 allées Antonio Machado, 31058 Toulouse cedex 9
prenom.nom@irit.fr

[2] LAAS, 7 Avenue du Colonel Roche, 31400 Toulouse
prenom.nom@laas.fr

Résumé : L'application **SWIP** (dont plusieurs aspects ont été présentés à la conférence IC ces dernières années) permet l'interrogation de base de connaissances à partir de requêtes exprimées en langage naturel. L'utilisation de patron permet à **SWIP** d'être cohérent lors de l'interprétation des questions. Ce papier de démo vise à présenter la dernière version de l'application qui simplifie son déploiement et rend son implémentation modulaire. Cette dernière version a été utilisée pour déployer **SWIP** sur une base de connaissances utilisée dans le cadre de l'accès aux données collectées par un batîment connecté.

Mots-clés : Interogation langage naturel, base de connaissances

1 Introduction

Les travaux menés depuis un demi-siècle sur la représentation de connaissances à l'aide de graphes étiquetés trouvent aujourd'hui une concrétisation à très grande échelle sur le Web Sémantique. Nous travaillons depuis quelques années au développement d'un environnement permettant d'interroger de manière aussi souple et aisée que possible les vastes entrepôts de données désormais disponibles dans les langages RDF (`https://www.w3.org/RDF/`, RDFS (`https://www.w3.org/TR/rdf-schema/`) et OWL (`https://www.w3.org/OWL/`) proposés par le W3C. Le Linking Open Data cloud diagram [1] donne un aperçu des entrepôts aujourd'hui disponibles dans ces langages. Le W3C a également standardisé le langage de requête SPARQL qui permet d'exprimer des requêtes dans une syntaxe "à la SQL". Formuler une requête en SPARQL implique de maîtriser le langage mais aussi les vocabulaires (ou schémas) avec lesquels les données ont été décrites dans l'entrepôt visé. La maîtrise de ce langage par un utilisateur final du Web est inenvisageable.

Le système **SWIP** (Semantic Web Interface using Patterns) proposé dans le cadre de la thèse de Pradel (2013), permet à l'utilisateur d'exprimer son besoin en information de manière intuitive sous forme de question en langage naturel. Nous présentons dans ce papier, une nouvelle version de ce système (en comparaison avec ce que nous avions présenté dans Pradel *et al.* (2013a)) qui repose maintenant sur un architecture modulaire facilitant son déploiement ainsi que sur une interface améliorant les interactions avec l'utilisateur.

1. `http://lod-cloud.net/`

2 Principe de SWIP

Afin de permettre une interprétation cohérente de la requête, **SWIP** repose sur un principe original (vis à vis des travaux existants présentés dans Höffner *et al.* (2016)) qui est la définition de patrons de requêtes. Ces patrons sont des modèles de requêtes que nous adaptons aux besoins exprimés par l'utilisateur. En effet, dans le cadre d'applications réelles, les requêtes soumises par les utilisateurs sont généralement des variations autour de quelques familles de requêtes typiques. L'utilisation de patrons, comme nous l'avons montré dans Pradel *et al.* (2013b), nous affranchit notamment de la phase d'exploration de l'ontologie en vue de lier les concepts identifiés à partir des mots-clés, puisque les relations potentiellement pertinentes apparaissent dans les patrons. Le processus bénéficie donc des familles de requêtes pertinentes pré-établies. A partir des évaluations présentées dans Pradel *et al.* (2013c), nous avons montré que l'utilisation de patrons est particulièrement pertinente pour l'interrogation d'une base de connaissances d'un domaine précis. **SWIP** est moins performant pour l'interrogation d'entrepôts généralistes tel que DBPedia.

Nous illustrons, dans ce papier, l'utilisation de **SWIP** à partir d'un cas d'étude réel portant sur l'interrogation d'une base de connaissances contenant les descriptions des capteurs et des données qu'ils collectent dans le cadre du bâtiment connecté ADREAM (`https://www.laas.fr/public/fr/le-projet-adream`). L'ontologie (IOT-O) utilisée pour représenter ces données est présentée dans Seydoux *et al.* (2016).

3 Architecture logicielle

Pour pouvoir facilement adapter **SWIP** à un entrepôt de données, nous proposons une nouvelle architecture du système. Le code source de l'application est disponible en ligne (`https://framagit.org/IRIT_UT2J/SWIP/SWIP2`).

L'architecture de **SWIP** repose sur des technologies permettant le déploiement rapide. MAVEN [2] est utilisé pour la gestion des dépendances. De cette manière le déploiement de **SWIP** sur une nouvelle machine peut se faire en quelques minutes. Par exemple, pour le déploiement de **SWIP** pour l'utilisation des données de ADREAM seulement 10 minutes ont été nécessaires (sans compter le temps de conception des patrons). Ceci est le temps nécessaire pour l'export des données et leur transformation sous le bon format de fichier, leur envoi sur le serveur et la configuration spécifique du projet "adream". La conception des patrons nécessite plus de temps puisqu'il faut connaître les familles de requêtes spécifiques à cette base de connaissances et le vocabulaire utilisé. L'interface décrite dans la section 4 permet de faciliter cette étape.

Lors du déploiement de **SWIP** il est nécessaire de spécifier les bases de connaissances qui seront interrogées. Pour cela un dossier spécifique pour chaque base de connaissances interrogable est créé. Lors de l'exécution de **SWIP** ces projets sont disponibles indépendemment grâce à une interface sous forme de services Web de type REST.

La modularité de l'architecture repose sur l'utilisation du framework OpenQA (Marx *et al.* (2014)) Ce framework permettant la définition de chaîne de processus générique est dédié au "Semantic Question Answering" (SQA). Nous l'avons utilisé afin de définir la chaîne de processus suivante :

2. https ://maven.apache.org

— Détection d'entités nommées
— Génération du graphe de dépendances syntaxiques (utilisation de Stanford CoreNLP [3])
— Génération de la requête pivot (format de requête intermédiaire permettant une représentation explicite des relations et non dépendante de la langue décrit dans Pradel (2013))
— Appariement de la requête aux patrons et calcul du score de confiance

Chaque étape peut être modifiée en définissant un nouveau processus dans la chaîne de traitements. Par exemple, il est facile de modifier le processus de génération du graphe de dépendances syntaxiques pour en utiliser un autre si jamais nous souhaitons utiliser une autre librairie que Stanford CoreNLP. Il est aussi possible de changer le processus de génération de la requête pivot pour l'adapter à une langue en particulier (ce processus utilisant le graphe de dépendances syntaxiques il est spécifique à une langue).

4 Interfaces

L'IHM permettant l'utilisation de **SWIP** se décompose en deux interfaces. La première, l'interface d'interrogation, permet de poser une question grâce à un champ texte et d'obtenir les résultats. La deuxième, l'interface d'administration, permet la visualisation des patrons et à terme proposera leur édition.

4.1 Interface d'interrogation

L'interface d'interrogation de **SWIP**, propose une barre de recherche où taper une requête en langage naturel. La figure 1 montre l'affichage proposé suite à la soumission de la requête "Which sensors observed the temperature of room H102". L'interface présente alors une reformulation de la demande, basée sur l'interprétation la plus pertinente trouvée par **SWIP** en fonction du patron utilisé. Ceci permet à l'utilisateur de vérifier que **SWIP** comprend bien sa demande. La réponse à la question (le résultat de la requête SPARQL) est directement proposée à l'utilisateur. Deux possibilités s'offrent à lui si l'interprétation de la requête ne le satisfait pas. Il peut soit sélectionner une autre interprétation suggérée par **SWIP** (utilisation d'un autre patron), soit spécifier ou généraliser des parties de sa question. Ce changement de degré de généralisation se fait par le biais de curseurs présents dans le détail de l'interprétation choisie : les différentes valeurs de généralisation sont affichées à côté du curseur correspondant, guidant ainsi l'utilisateur dans la reformulation de sa requête. La figure 1 illustre la généralisation du mot "sensors" de la requête initiale "a physical object". Cette généralisation utilise la hiérarchie des classes présente dans le vocabulaire interrogé.

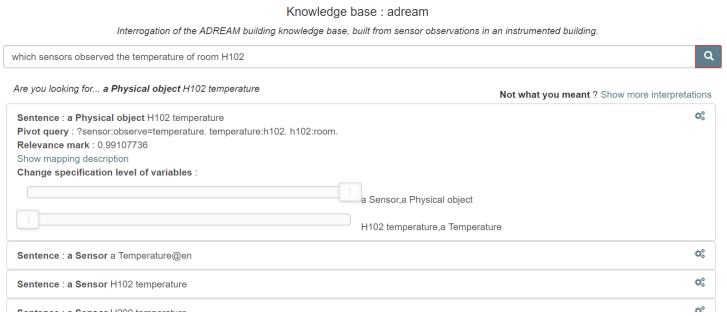


FIGURE 1 – Exemple de requête généralisée

3. https://stanfordnlp.github.io/CoreNLP/

4.2 Interface d'administration

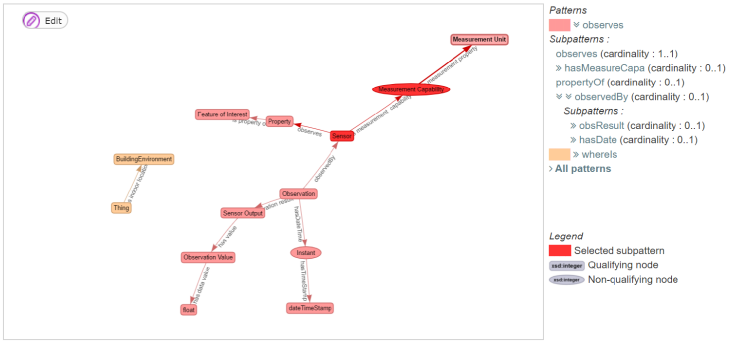


FIGURE 2 – Affichage d'un patron

SWIP comporte également une interface d'administration, permettant la visualisation de patrons de requêtes. La hiérarchie des patrons et sous-patrons est disponible dans la colonne de droite (c.f. figure 2). Au survol du nom d'un patron, ce dernier est mis en évidence par une couleur de surbrillance, tandis que les autres noeuds du graphe sont grisés. Cela permet une rapide identification des différents patrons, rendant aisée la découverte des différentes façon d'interroger la base de connaissances. La figure 2 illustre la hiérarchie partiellement déroulée, ainsi que l'affichage du graphe lors d'un clic sur le sous-patron *hasMeasureCapability*.

5 Travaux futurs

Nous souhaitons continuer ces travaux en facilitant la création et l'édition de patrons grâce à l'interface d'administration. L'affichage des patrons sous forme de graphe permet une modification et une création plus intuitive pour l'utilisateur que le système actuel (syntaxe spécifique).

Références

HÖFFNER K., WALTER S., MARX E., USBECK R., LEHMANN J. & NGONGA NGOMO A.-C. (2016). Survey on challenges of question answering in the semantic web. *Semantic Web*, (Preprint), 1–26.

MARX E., USBECK R., NGOMO A.-C. N., HÖFFNER K., LEHMANN J. & AUER S. (2014). Towards an open question answering architecture. In *Proceedings of the 10th International Conference on Semantic Systems*, p. 57–60 : ACM.

PRADEL C. (2013). *D'un langage de haut niveau à des requêtes graphes permettant d'interroger le web sémantique*. Thèse de doctorat, Université de Toulouse, Toulouse, France. (Soutenance le 05/12/2013).

PRADEL C., HAEMMERLÉ O. & HERNANDEZ N. (2013a). Demo : Swip, a semantic web interface using patterns. In *Proceedings of the ISWC 2013 Posters & Demonstrations Track, Sydney, Australia, October 23, 2013*, p. 73–76.

PRADEL C., HAEMMERLÉ O. & HERNANDEZ N. (2013b). Passage de la langue naturelle à une requête SPARQL dans le système SWIP. In *IC 2013 : 24es Journées francophones d'Ingénierie des Connaissances (Proceedings of the 24th French Knowledge Engineering Conference), Lille, France, July 1-5, 2013.*, p. 207–222.

PRADEL C., PEYET G., HAEMMERLÉ O. & HERNANDEZ N. (2013c). Swip at qald-3 : results, criticisms and lesson learned.

SEYDOUX N., DRIRA K., HERNANDEZ N. & MONTEIL T. (2016). IoT-O, a core-domain IoT ontology to represent connected devices networks. In *EKAW*.

Le projet ECOPACK : environnement socio-technique support à une analyse stratégique

Claude Moulin[1], Marie-Hélène Abel[1], Véronique Misséri[2]

[1] Sorbonne universités, Université de technologie de Compiègne
UMR CNRS 7253, LABORATOIRE HEUDIASYC, CS 60319,
60203 Compiègne Cedex, France,
{claude.moulin, marie-helene.abel}@utc.fr

[2] Sorbonne universités, Université de technologie de Compiègne
EA 2223 COSTECH CS 60319,
60203 Compiègne Cedex, France,
veronique.misseri@utc.fr

Résumé : Dans le cadre du projet ECOPACK nous nous intéressons aux nouvelles formes de travail collaboratif permises par un environnement socio-technique informatique incluant différents dispositifs (table, tableau et tablette tactiles, smartphone, pc). Notre objectif principal est de définir un écosystème numérique capable de répondre aux besoins d'idéation, d'innovation et d'analyse stratégique. La plateforme ECOPACK a été développée et testée afin de répondre aux besoins de collaboration identifiés dans le cadre d'une analyse stratégique autour d'une carte/graphe. La démonstration de l'application ECOPACK se fera à partir des données capitalisées pour une analyse stratégique concernant les enjeux de la mer.

Mots-clés : plateforme de collaboration, écosystème de connaissances, capitalisation des connaissances

1 Contexte

Dans le cadre du projet ECOPACK Saleh *et al.* (2016b) nous nous intéressons aux nouvelles formes de travail collaboratif permises par un environnement socio-technique informatique incluant différents dispositifs (table, tableau et tablette tactiles, smartphone, pc).

Un des objectifs est d'offrir une application partagée accessible à chaque collaborateur via ces différents dispositifs. Chaque dispositif privilégie un type d'interaction ainsi qu'un type d'activité. Les participants devront pouvoir exploiter leurs propres ressources pour collaborer. L'application partagée doit donc donner accès à un système d'information plus large ou chaque participant est utilisateur et chaque groupe de collaboration est identifié. L'utilisation collaborative de l'application partagée va elle-même produire des ressources spécifiques qu'il s'agit de capitaliser au même titre que toute autre ressource dans le système d'information. Capitaliser les traces d'interactions est également souhaitable afin de les exploiter dans des justifications de prises de décision. L'approche suivie rejoint la vision d'écosystème de connaissances qui favorise l'évolution dynamique des interactions de connaissances entre des collaborateurs afin d'améliorer la prise de décision et l'innovation Chesbrough (2006) Nous visons particulièrement celle des écosystèmes d'affaire Dini *et al.* (2005).

Notre objectif principal est donc de définir un écosystème numérique capable de répondre aux besoins d'idéation, d'innovation et d'analyse stratégique. Nous avons défini un modèle de plateforme intégrant les dispositifs adaptés à des liens forts entre les acteurs pour augmenter les performances de l'écosystème. Nous visons à permettre à chaque collaborateur de contribuer

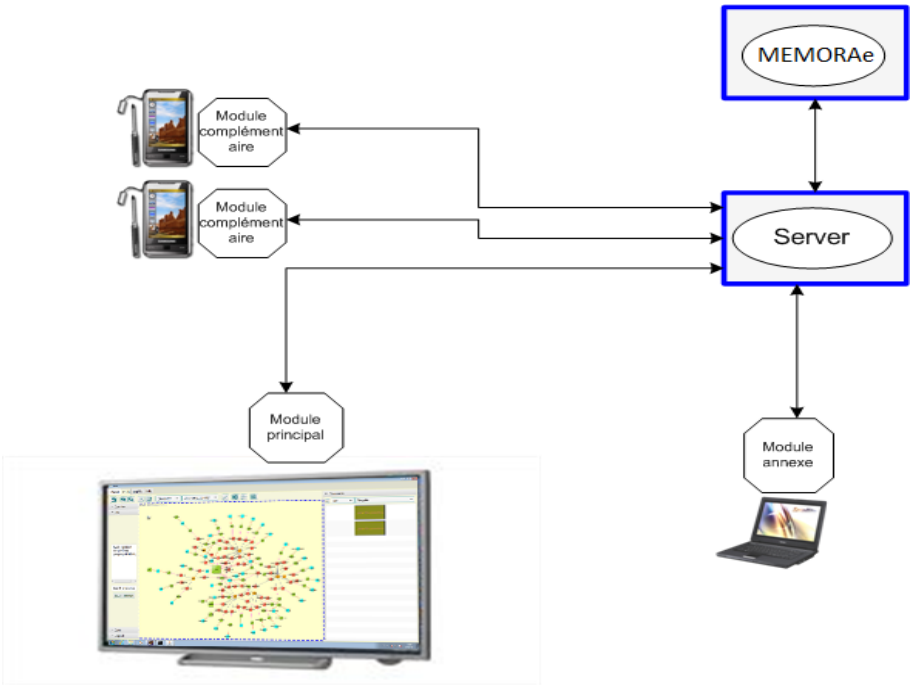


FIGURE 1 – Architecture de la plateforme ECOPACK.

à l'écosystème à tout moment selon les dispositifs à disposition. Cela nécessite un système d'information multi-utilisateur multi-support.

Les ressources créées et partagées par l'écosystème requièrent à la fois un travail individuel et collaboratif. La plateforme élaborée a pour vocation de permettre une continuité spatio-temporelle des séances de travail. Un utilisateur doit pouvoir préparer individuellement une prochaine séance collaborative.

2 Plateforme

La plateforme ECOPACK a été développée et testée afin de répondre aux besoins de collaboration identifiés dans le cadre d'une analyse stratégique autour d'une carte/graphe Saleh *et al.* (2016a). Il s'agit d'une application distribuée comprenant plusieurs modules (Fig. 1) :

— Un serveur en liaison avec une plateforme de collaboration (MEMORAe-ECOPACK) qui permettra de capitaliser des connaissances autour des données gérées par application ECOPACK.
— Un module principal destiné à un grand tableau tactile.
— Des modules complémentaires destinés aux tablettes tactiles.
— Un module annexe donnant une représentation adaptée de celle du module principal et destiné à des PCs.

Un projet ECOPACK est un ensemble de données hétérogènes regroupées en items, présentées et recueillies par le module principal.

— Le module principal est chargé d'offrir une représentation dynamique de ces données qui en permet l'interprétation par des experts (Fig. 2).

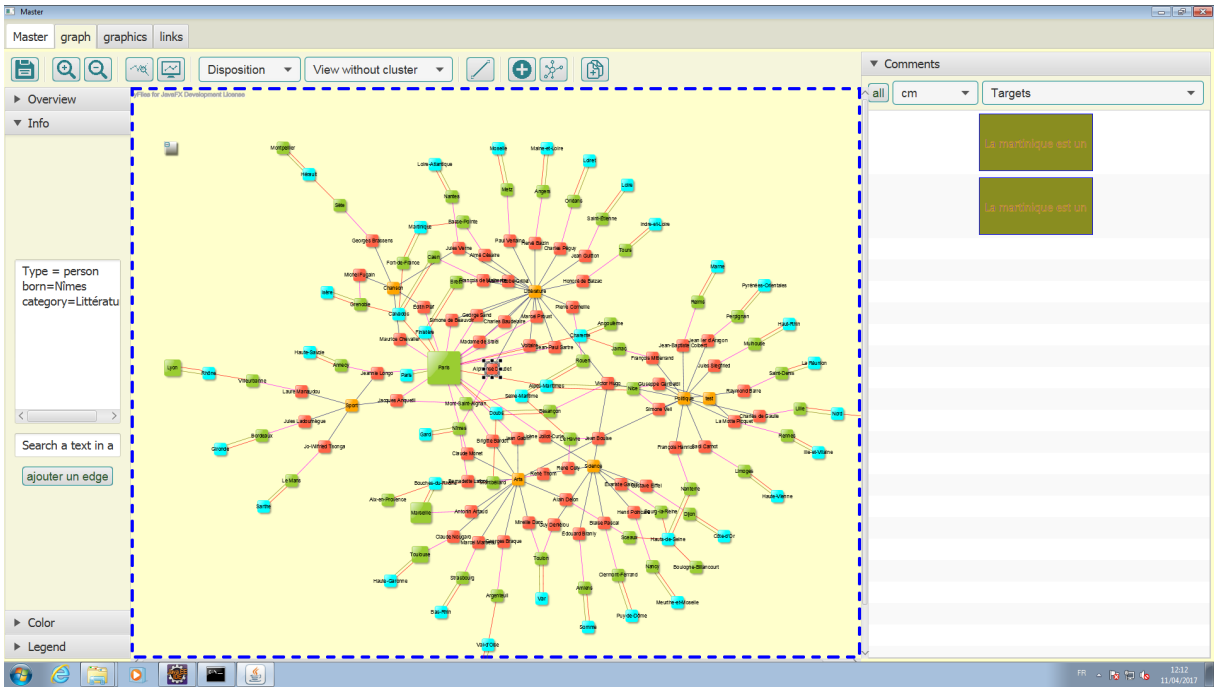


FIGURE 2 – Interface du module principal.

— Les modules complémentaires permettent aux participants d'une session de travail d'envoyer des commentaires au module principal concernant les données affichées.
— Le module annexe offre une représentation sur une surface réduite des informations visibles sur le module principal.

3 Démonstration

La démonstration de la plateforme ECOPACK se fera à partir des données capitalisées pour une analyse stratégique concernant les enjeux de la mer. Il s'agissait de pouvoir répondre à la question suivante proposée par l'association régionale AR24 Picardie de l'Institut des Hautes Etudes de Défense Nationale (IHEDN) : L'industrie navale française, entre stratégie d'innovation technologique et d'influence, quels atouts faire valoir pour renforcer la position de la France sur l'échiquier mondial ?

Pour ce faire, dans le cadre de l'enseignement d'Intelligence Economique EI04 dispensé à l'Université de Technologie de Compiègne (UTC), quatre groupes d'étudiants sur un an ont bâti une base de données riche et fiable sur les relations entre d'une part les entreprises, fournisseurs, acheteurs et financeurs de matériels navals militaires et d'autre part les conflits armés et les rencontres diplomatiques. Une séance d'analyse stratégique a été conduite à partir de cette base présentée sous forme de carte. La carte a servi de support à l'émission d'hypothèses et de graphiques. L'animation a impliqué la manipulation des données au travers de la carte par des focus et des filtres. Elle a été conduite avec le concours de 5 experts qui ont émis par le biais des modules complémentaires des commentaires, des questionnements, tout au long de la séance. Ces interventions ont été enregistrées et indexées comme une ressource à part entière pour des

séances futures afin de dynamiser le processus écosystémique qui pourra se prolonger au-delà de la séance.

4 Remerciements

Ce travail a été réalisé dans le cadre du projet ANR ECOPACK (ANR-13-ASTR-0026) sur l'Accompagnement Spécifique des Travaux de Recherches et d'Innovation Défense (ASTRID).

Références

CHESBROUGH H. W. (2006). *Open innovation : The new imperative for creating and profiting from technology*. Harvard Business Press.

DINI P., DARKING M., RATHBONE N., VIDAL M., HERNANDEZ P., FERRONATO P., BRISCOE G. & HENDRYX S. (2005). The digital ecosystems research vision : 2010 and beyond. *European Commisssion, Bruxelles, Position Paper*.

SALEH M., ABEL M.-H., MISSÉRI V., MOULIN C. & VERSAILLES D. (2016a). Integration of brainstorming platform in a system of information systems. In *Proceedings of the 8th International Conference on Management of Digital EcoSystems*, p. 166–173 : ACM.

SALEH M., MISSÉRI V. & ABEL M.-H. (2016b). Managing heterogeneous information in a system of information systems. In *KMIS 2016 8th International Conference on Knowledge Management and Information Sharing*.

Système de systèmes d'information et écosystème apprenant

Majd Saleh[1], Mohamed Ali Ben Ameur[1], Marie-Hélène Abel[1]

Sorbonne universités, Université de technologie de Compiègne
UMR CNRS 7253, LABORATOIRE HEUDIASYC, CS 60319,
60203 Compiègne Cedex, France,
{majd.salah, mohamed-ali.ben-ameur, marie-helene.abel}@utc.fr

Résumé : Avec le développement des technologies de l'information et de la communication, les apprenants sont confrontées à une grande quantité d'informations issues de nombreux systèmes. Gerer les ressources hétérogènes d'un écosystème apprenant devient un vrai challenge. A cette fin, nous avons fait le choix de modéliser un écosystème apprenant comme un système de systèmes d'information (SoIS) au sein duquel nous introduisons un système de vote et d'annotation de ressources. Dans notre SoIS, nous tenons compte de qui recommande quoi à qui sur quel sujet, quand, comment et pourquoi. Nous le faisons au moyen du modèle de collaboration mis en œuvre pour la conception du SoIS support à l'écosystème apprenant.

Mots-clés : système de systèmes d'information, écosystèmes apprenants, système de vote, système de recommandation.

1 Contexte

Dans le contexte de « l'apprentissage ensemble », de nombreux systèmes d'information sont utilisés par les apprenants afin d'exploiter des ressources hétérogènes (vidéo, texte, forum en ligne, etc.). Ils doivent alors s'organiser afin d'accéder aux ressources distribuées dans ces systèmes de façon à savoir où se trouvent les ressources pertinentes sur un sujet donné. Cette organisation est essentielle dans la mesure où elle vise à permettre l'accès à la bonne information au bon moment. Les apprenants évoluent ainsi dans un écosystème apprenant comprenant les apprenants eux-mêmes et leur environnement physique et social.

A l'ère des technologies 2.0, les écosystèmes numériques ont pour objectif de garantir le partage des connaissances au sein des organisations aussi rapidement et efficacement que possible Price & Turnbull (2007). Ils peuvent être considérés comme des plateformes support à la coopération, au partage et à l'accès aux connaissances afin de faciliter l'apprentissageAgostinelli & Koulayan (2016).

De leur côté, les systèmes de systèmes (SoS) sont des collections de systèmes dédiés qui regroupent leurs ressources et leurs capacités pour créer un nouveau système plus complexe qui offre plus de fonctionnalités que simplement la somme des systèmes composants Popper *et al.* (2004). Dans notre contexte, nous nous intéressons plus particulièrement à une catégorie de SoS : les systèmes de systèmes d'information (SoIS). Un SoIS peut être considéré comme un macro-système d'information donnant accès aux informations distribuées dans les systèmes composants et offrant des fonctionnalités utilisant les informations accédées.

Afin de faciliter l'organisation des ressources au sein d'un écosystème apprenant numérique, nous avons fait le choix de le considérer comme un système de systèmes d'information et d'en développer une plateforme support à partir d'un modèle de collaboration.

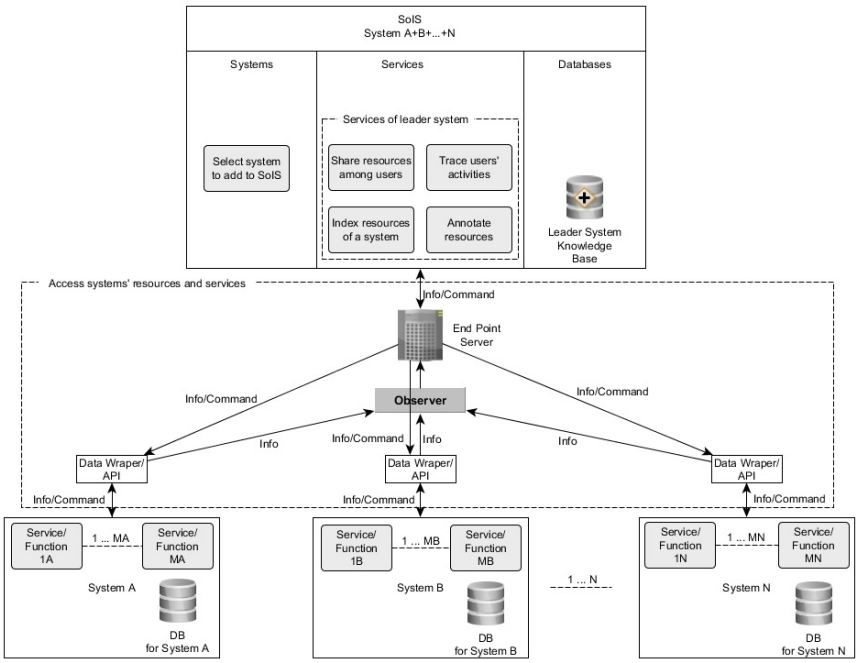


FIGURE 1 – Architecture du MEMORAeSoIS. Saleh & Abel (2016)

2 Plateforme

La plateforme MEMORAeSoIS a été développée à partir d'une architecture de SoIS et du modèle de collaboration memorae-core2 Abel (2015). Ce modèle vise à faciliter la gestion des ressources d'information hétérogènes au sein des organisations et l'apprentissage organisationnel.

Le modèle memorae-core2 est une ontologie modulaire reprenant des standards du Web sémantique : FOAF (Friend Of A Friend) Brickley & Miller (2010), SIOC (Semantically-Interlinked Online Communities) Breslin *et al.* (2009), BIBO (BIBliographic Ontology) D'Arcus & Giasson (2009), OA (Open Annotation). La collaboration est considérée du point de vue du partage et de l'échange de ressources hétérogènes de connaissances entre collaborateurs utilisateurs. Le coeur du modèle s'organise naturellement entre les concepts utilisateurs, groupe d'utilisateurs et ressources Atrash *et al.* (2014).

L'architecture de MEMORAeSoIS (Fig. 1) suit l'approche Follower/Leader : un système leader permet aux systèmes composants de coopérer, de mener une tâche en collaboration Dong & Hussain (2007).

Le système leader de MEMORAe SoIS comporte un groupe de connecteurs de systèmes et de bases de données comprenant les informations (identifiant/mot de passe) liées aux connections aux systèmes composants.

Développé à partir de memorae-core2 le système leader du MEMORAeSoIS propose aux utilisateurs apprenants les fonctionnalités suivantes :

— Accéder aux ressources provenant de différents systèmes d'information dans un emplacement centralisé,

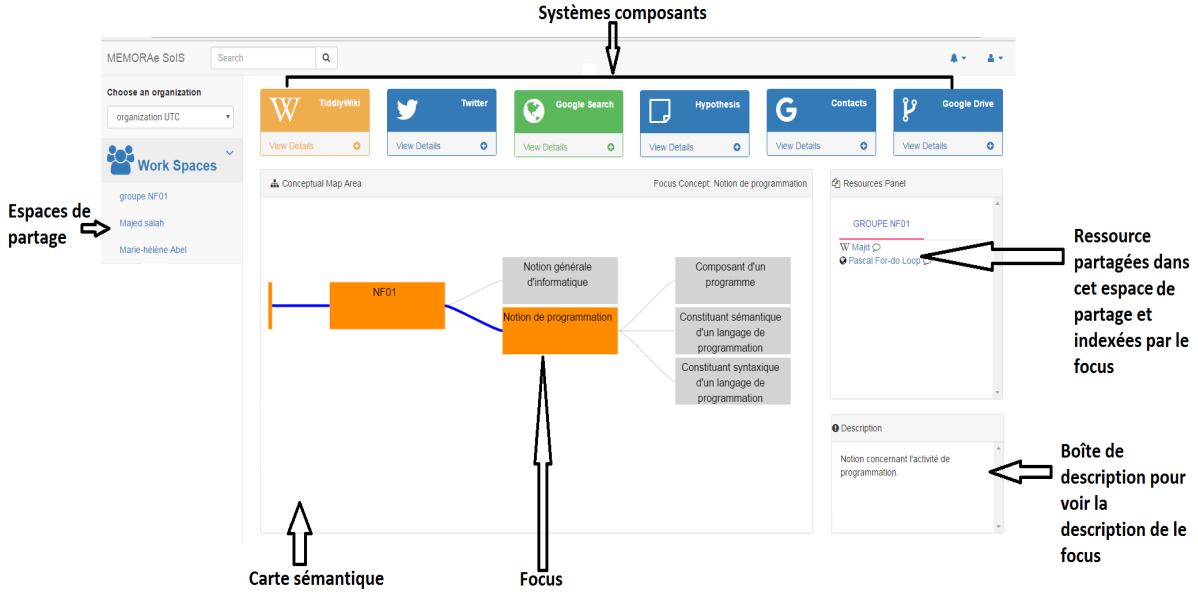


FIGURE 2 – Ecran principal du MEMORAeSoIS

— Organiser/indexer les ressources autour d'un référentiel partagé présenté sous la forme d'une carte (ontologie d'application),
— Partager les ressources dans différents espaces de partage,
— Annoter/Voter les ressources afin de mettre en évidence certaines idées liées aux ressources.

3 Démonstration

Actuellement, nous travaillons avec les étudiants de l'Université de Technologie de Compiègne (UTC) pour évaluer la plateforme MEMORAeSoIS. Les étudiants sont inscrits en cours d'algorithmique (NF01). Pour la configuration de ce test, chaque étudiant reçoit un compte dans le SoIS. Ce compte donne accès à une carte de concepts définissant le contenu du cours (notion à appréhender/référentiel partagé) de l'organisation NF01. Les étudiants peuvent ensuite organiser, partager via le système leader les ressources issues des systèmes composants. Au sein du système leader, chaque étudiant dispose de trois espaces de partage dans la plateforme : un espace de partage personnel, un espace de partage pour le groupe de travail et un dernier espace de partage avec l'enseignant du cours. Les ressources sont indexées par une ou plusieurs notions définies dans la carte de concepts.

Les systèmes composants de MEMORAeSoIS (Fig. 2) sont TiddlyWiki (gestion de contenu de page wiki), Twitter, Google Search, hypothesis (annotation de pages web), Google Contacts, Google Drive (stockage de fichiers).

Le système leader propose une fonctionnalité de vote participatif (Fig. 3) aux apprenants afin qu'ils puissent exprimer leur point de vue sur la pertinence d'une ressource pour appréhender le concept qui l'indexe. Un degré de pertinence sera alors calculé en tenant compte de l'ensemble des votes au sein d'un espace de partage. Un étudiant vote pour une ressource concernant un

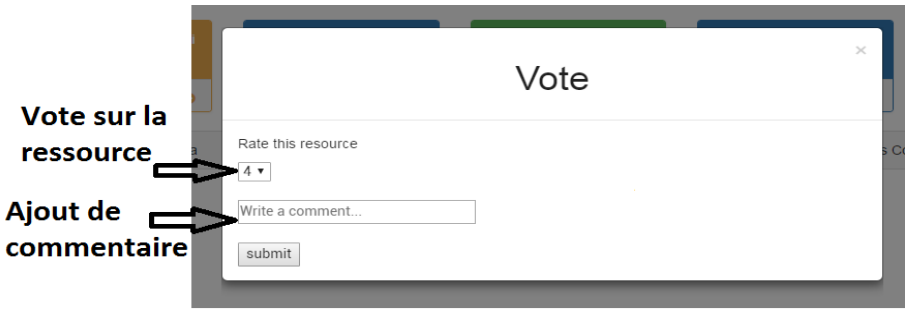


FIGURE 3 – Fonctionnalité de vote

sujet pour un public cible.

4 Conclusion

Dans le cadre de notre travail nous avons fait le choix de modéliser un écosystème apprenant comme un système de systèmes d'information (SoIS) au sein duquel nous introduisons un système de vote des utilisateurs de l'écosystème. Dans notre approche, nous tenons compte de la volonté de collaborer. Par conséquent, nous avons développer la plateforme MEMORAeSoIS à partir du modèle de collaboration memorae-core2 afin de pouvoir développer des fonctionnalités exploitant les informations comme qui collabore avec qui, comment, quand, pourquoi, sur quoi et où.

Références

ABEL M.-H. (2015). Knowledge map-based web platform to facilitate organizational learning return of experiences. *Computers in Human Behavior*, **51**, 960–966.

AGOSTINELLI S. & KOULAYAN N. (2016). Les écosystèmes numériques.

ATRASH A., ABEL M.-H. & MOULIN C. (2014). Supporting organizational learning with collaborative annotation. In *International Conference on Knowledge Management and Information Sharing*, p. 237–244.

BRESLIN J., PASSANT A. & DECKER S. (2009). *The social semantic web*. Springer Science & Business Media.

BRICKLEY D. & MILLER L. (2010). Foaf vocabulary specification 0.98. namespace document 9 august 2010-marco polo edition.

D'ARCUS B. & GIASSON F. (2009). Bibliographic ontology specification. specification document, 4 november 2009. *Retrieved August*, **10**, 2011.

DONG H. & HUSSAIN F. K. (2007). Digital ecosystem ontology. In *Industrial Electronics, 2007. ISIE 2007. IEEE International Symposium on*, p. 2944–2947 : IEEE.

POPPER S. W., BANKES S. C., CALLAWAY R. & DELAURENTIS D. (2004). System of systems symposium : Report on a summer conversation. *Potomac Institute for Policy Studies, Arlington, VA*, **320**.

PRICE C. & TURNBULL D. (2007). The organizational challenges of global trends : A mckinsey global survey. *McKinsey Quarterly*.

SALEH M. & ABEL M.-H. (2016). Moving from digital ecosystem to system of information systems. In *Computer Supported Cooperative Work in Design (CSCWD), 2016 IEEE 20th International Conference on*, p. 91–96 : IEEE.

Un outil de catégorisation conceptuelle des formations professionnelles

Mohamed Nader Jelassi[1,2], Sylvie Ranwez[1]
Sébastien Harispe[1], Jacky Montmain[1], Christophe Blondeau[2]

[1] LGI2P, École des mines d'Alès, 69 rue Georges Besse F-30035 Nîmes cedex 1
prenom.nom@mines-ales.fr.

[2] Edotplus SAS, 442 rue Georges Besse 30035 Nîmes Cedex 1
christophe.blondeau@edotplus.com
http://www.edotplus.fr

Résumé : Afin d'aider les formateurs à optimiser le taux de remplissage de leurs salles, et les salariés ou demandeurs d'emplois à bénéficier de parcours de formation pertinents pour répondre à leur réorientation ou évolution de carrière, nous proposons une solution logicielle innovante basée sur une approche sémantique de la gestion des formations. Cette solution repose sur une ontologie des formations professionnelles qui supporte la classification des formations ainsi que les calculs permettant la recommandation. La présente démonstration concerne l'ontologie et son utilisation pour la classification des formations.

Mots-clés : Ontologie de domaine, Indexation conceptuelle, Recommandation, Formation professionnelle

1 Introduction et Objectifs

Edotplus [1] est une jeune start-up qui se positionne comme médiateur entre des sociétés de formation et des demandeurs de formation. En coopération avec l'École des mines d'Alès, elle propose de mettre en place un SaaS (software as a service) basé sur une approche sémantique de la gestion des formations. La solution proposée est intégrée à une application qui vise à aider les formateurs à optimiser le taux de remplissage de leurs salles. Dans le même temps, cette application recommande des parcours de formation pour des salariés ou des demandeurs d'emplois qui ont un projet d'évolution de carrière (*e.g.*, réorientation, adaptation à un nouveau poste). Cette recommandation tient compte de leurs profils (diplômes, acquis professionnels) et propose un parcours au meilleur rapport qualité/prix, pour une formation tout au long de leur vie.

Trois axes de recherche et développement ont été identifiés :

— La définition d'une ontologie du domaine de la formation professionnelle qui doit servir de base aux différents calculs sémantiques liés à l'application ;
— L'indexation conceptuelle des formations (classification) ;
— La mise en place de traitements pour la recommandation : regroupement en fonction de certains critères (clustering), recherche d'information conceptuelle.

La démonstration qui fait l'objet du présent article concerne uniquement les deux premiers points. La section suivante présente l'ontologie construite et son évolution. L'approche proposée pour la classification est présentée dans la section 3. Enfin la section 4 discute les résultats obtenus lors des premières évaluations.

1. http ://www.e.pluswww.e.plus

2 Ontologie de la formation professionnelle

L'application d'une approche conceptuelle pour la recherche d'information (Sy *et al.*, 2012) ou pour l'indexation conceptuelle (Fiorini *et al.*, 2015) impose de disposer d'une modélisation de la connaissance du domaine (*e.g.* ontologie). Les différentes ontologies pédagogiques de la littérature étant plus orientées vers les techniques pédagogiques, elles n'étaient pas adaptées à notre contexte. Pour construire cette ontologie, nous avons répertorié puis organisé les différents critères communément admis pour décrire une formation, *e.g.,* sa durée, son public cible, ses modalités, ou encore son coût. Cependant la partie centrale de l'ontologie concerne les différentes disciplines auxquelles sont rattachées les formations. Une organisation hiérarchique de ces disciplines a dû être définie (*e.g.*, Java est une spécialisation de Langage de programmation, lui-même spécialisation d'Informatique). C'est cette hiérarchie qui permettra notamment, par la suite, d'estimer des distances sémantiques entre formations rattachées à ces différents concepts. Pour obtenir cette organisation hiérarchique, différentes classifications non-standardisées utilisées par les organismes de formation ont été unifiées. Nous avons rencontré quelques difficultés dans la modélisation ayant pour origine : i) les classifications multiples produites par les experts mais parfois insuffisamment précises, ii) le multilinguisme et certaines ambiguïtés de langages (e.g., Finance en français réfère à des activités bancaires tandis que le même terme en anglais réfère à du contrôle de gestion), iii) Certains titres de formations peu explicites (e.g., CACES R389 qui est une formation de conduite d'engins) et ne peuvent pas aider à la classification des formations. Il a été nécessaire de faire appel à des experts pour résoudre certains conflits ou parvenir à la traduction la plus adaptée. L'ontologie ainsi obtenue contient 997 concepts, chacun étant associé à un label en français et un label en anglais.

Un premier modèle opérationnel a été mis en place. Cette ontologie n'est certes pas définitive. Elle est amenée à évoluer en fonction des retours qui seront faits suite à son utilisation dans le système. Elle constitue cependant une base suffisante pour poursuivre le développement de la chaîne de traitement.

3 Classification des formations

A chaque formation est associée une indexation conceptuelle. En effet, en comparant les labels des différentes disciplines de l'ontologie avec le titre et/ou la description textuelle de chaque formation, un lien sémantique est établi entre ces formations et les disciplines auxquelles elles peuvent être rattachées. Une première classification a été réalisée sur un ensemble restreint de formations à partir d'analyse de texte : comparaison des différents labels des concepts avec les titres des formations en utilisant la distance de Levenshtein (Levenshtein, 1999). Cette classification a été confrontée à des avis d'experts et l'association entre les différentes formations et les disciplines de l'ontologie a été ainsi validées. Cette base d'apprentissage est ensuite utilisée pour catégoriser l'ensemble des formations de la plate-forme (plusieurs dizaines de milliers). Pour ce faire, une distance lexicale est calculée entre chaque description d'une nouvelle formation à indexer et l'ensemble des descriptions des formations de la base d'apprentissage (déjà indexées). L'indexation des formations les plus proches au sens de cette distance est reportée comme indexation de cette formation considérée. Lors de l'exploitation de ce modèle, ce processus sera appliqué pour chaque nouvelle formation saisie par un organisme sur son espace dédié. Dans ce contexte, l'outil proposera un ensemble de catégories à l'utilisateur qui pourra

k	Précision
1	0.41
2	0.57
3	0.68
4	0.75
5	0.79

TABLE 1 – Précision de notre outil de recommandation pour différentes valeurs de k.

valider ou non la classification proposée pour sa formation.

Cette classification unifiée favorise l'application d'un même traitement et permet de proposer, dans un même outil, des recommandations de formations issues de différents organismes.

4 Évaluation de l'ontologie et de la classification des formations

Base d'apprentissage/Base de test et protocole de validation

Afin de réaliser nos expérimentations et d'évaluer la pertinence de notre outil, nous avons utilisé un protocole de "validation croisée" (ou cross-validation) (Weiss & Kulikowski, 1991). L'ensemble de formations de notre ontologie (27107 formations au total dans différentes langues [2]) a été partitionné en deux échantillons : un échantillon aléatoire contenant 80% des formations a été utilisé comme base d'apprentissage et l'échantillon contenant les 20% de formations restantes, a été utilisé pour la validation de nos tests (*i.e.,* base de test). Pour chaque formation, nous proposons une liste de disciplines qui peuvent lui être associées. Si dans cette liste figure une discipline initialement donnée par l'expert, alors elle est considérée comme pertinente pour l'indexation. Pour nos expérimentations, nous avons également fait varier la taille de la liste de disciplines suggérées pour l'indexation. Il s'agit des top-k recommandations. Ainsi, l'utilisateur peut spécifier les k disciplines les plus pertinentes que le système doit lui retourner. Les k premières réponses sont celles qui ont les scores (la moyenne des distances de Levenshtein) les plus élevés. Et afin de déterminer l'efficacité de notre outil, nous appliquons une métrique classique en recherche d'information : la précision (Baeza-Yates & Ribeiro-Neto, 1999).

Le Tableau 1 montre les différentes valeurs de précision obtenues par notre outil de recommandation pour différentes valeurs de k variant entre 1 et 5. En général, nos recommandations pour les formateurs correspondent aux catégories attendues, *i.e.,* celles affectées par les experts. En effet, les recommandations sont pertinentes jusqu'à 79% pour k=5. Notre outil de classification permet en outre de limiter les saisies en offrant à l'utilisateur des mots-clés pour classer sa formation. De plus, cela permet d'enrichir directement le contenu d'une formation, *i.e.,* une formation appartenant à une même catégorie qu'une autre formation pourra enrichir ses caractéristiques (*e.g.,* public visé, description,...).

2. les formations sont issues de 6 sites européens proposant des offres de formations.

5 Conclusion et Perspectives

Dans ce papier, nous avons présenté notre outil pour la catégorisation de formations professionnelles. Les premiers résultats démontrent un très bon niveau de classification automatique. Cependant, ils peuvent être améliorés en exploitant d'autres critères de formations comme la description du contenu ou encore le public visé pour cibler au mieux la catégorie à laquelle la formation est susceptible d'appartenir. De plus, nous souhaitons proposer des recommandations de formations professionnelles appartenant à des catégories similaires. Enfin, le besoin d'uniformiser diverses sources sur une même classification va nous permettre d'analyser des comportements de cohortes (prix, densité concurrentielle,...).

Références

BAEZA-YATES R. A. & RIBEIRO-NETO B. (1999). *Modern Information Retrieval*. Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc.

FIORINI N., RANWEZ S., MONTMAIN J. & RANWEZ V. (2015). Usi : a fast and accurate approach for conceptual document annotation. *BMC Bioinformatics*, **16**(1), 83.

LEVENSHTEIN V. I. (1999). Equivalence of delsarte's bounds for codes and designs in symmetric association schemes, and some applications. *Discrete Mathematics*, **197**, 515 – 536.

SY M.-F., RANWEZ S., MONTMAIN J., REGNAULT A., CRAMPES M. & RANWEZ V. (2012). User centered and ontology based information retrieval system for life sciences. *BMC Bioinformatics*, **13**(1), S4.

WEISS S. M. & KULIKOWSKI C. A. (1991). *Computer Systems That Learn : Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.