Actes des 30es journées francophones d'Ingénierie des Connaissances

Nathalie Jane Hernandez

AfIA

Association française
pour l'Intelligence Artificielle

IC

Journées francophones d'Ingénierie des Connaissances

PFIA 2019

Table des matières

Éditorial

Les journées francophones d'Ingénierie des Connaissances sont organisées chaque année depuis 1997. Elles le furent tout d'abord sous l'égide du Gracq (Groupe de Recherche en Acquisition des Connaissances) et le sont aujourd'hui sous celle du collège SIC (Science de l'Ingénierie des Connaissances) de l'AFIA.

L'ingénierie des connaissances peut être vue comme la partie de l'Intelligence Artificielle se préoccupant des connaissances selon les points de vue de la représentation, l'acquisition et l'intégration dans des environnements numériques. Sa finalité est la production de méthodes et outils "intelligents", capables d'aider l'humain dans ses activités.

La conférence Ingénierie des Connaissances réunit la communauté francophone et est un lieu d'échanges et de réflexions, de présentation et de confrontation des théories, pratiques, méthodes et outils. Cette communauté doit désormais prendre en compte l'essor des algorithmes d'apprentissage et leurs retombées sur les pratiques individuelles et collectives.

Un autre des enjeux récents est la préservation de la vie privée, dans les approches d'IC, qui collectent et traitent des données personnelles. La conférence invitée de **Ruben Verborgh** intitulée **"How Solid aims to impact the Web (and AI with it)"** porte sur ce thème.

Pour cette édition de la conférence, 43 articles ont été soumis. 24 d'entre-eux ont été acceptés pour présentation longue et 10 pour présentation à partir de posters.

Parmi les papiers acceptés, 6 portent sur des travaux académiques originaux et 3 sur des travaux menés en collaboration avec des entreprises. Cette édition est également l'occasion de proposer une rétrospective des travaux passés et de prospecter les directions que pourrait prendre l'IC à l'avenir. Dans cette perspective, 11 papiers correspondent à des travaux marquants de l'IC déjà publiés en version longue à l'international. Ces derniers sont présentés dans ces actes sous la forme de résumés. De plus, 4 papiers présentent des positionnements en lien avec les avancées de l'IC de ces dernières années.

Nathalie Hernandez

Comité de programme

Présidents

— Nathalie Hernandez IRIT Université Toulouse 2 Jean-Jaurès

Membres

— Marie-Hélène Abel UTC
— Xavier Aimé Cogsonomy / LIMICS UMRS 1142 Inserm
— Yamine Ait Ameur IRIT/INPT-ENSEEIHT
— Bruno Bachimont Sorbonne Université
— Jean-Paul Barthès UTC
— Aurélien Bénel Université de technologie de Troyes
— Nacéra Bennacer Seghouani LRI CentraleSupélec
— Bertrand Braunschweig INRIA
— Nathalie Bricon-Souf IRIT Université Paul Sabatier Toulouse
— Sandra Bringay LIRMM
— Patrice Buche INRA
— Davide Buscaldi LIPN, Université Paris 13, Sorbonne Paris Cité
— Sylvie Calabretto LIRIS
— Gaoussou Camara Université Alioune Diop de Bambey - Sénégal
— Pierre-Antoine Champin LIRIS, Université Claude Bernard Lyon1
— Jean Charlet AP-HP & INSERM UMRS 1142
— Olivier Corby INRIA
— Mathieu D'Aquin Insight Centre for Data Analytics, National University of Ireland Galway
— Sylvie Despres Laboratoire d'Informatique Médicale et de BIOinformatique (LIM&BIO)
— Jean-Pierre Evain EBU
— Gilles Falquet University of Geneva
— Catherine Faron Zucker Université Nice Sophia Antipolis
— Cécile Favre ERIC - Université Lyon 2
— Béatrice Fuchs LIRIS, IAE - université Lyon 3
— Frédéric Fürst MIS - Université de Picardie - Jules Verne
— Alban Gaignard CNRS
— Jean-Gabriel Ganascia Pierre and Marie Curie University - LIP6
— Serge Garlatti IMT Atlantique
— Alain Giboin INRIA
— Nathalie Guin LIRIS - Université de Lyon
— Mounira Harzallah LS2N
— Ollivier Haemmerlé IRIT
— Liliana Ibanescu AgroParisTech, INRA,
— Sébastien Iksal LIUM - Le Mans Université, France
— Antoine Isaac Europeana & VU University Amsterdam
— Clement Jonquet University of Montpellier - LIRMM
— Mouna Kamel IRIT - Université Paul Sabatier - Toulouse
— Gilles Kassel University of Picardie Jules Verne
— Pascale Kuntz Laboratoire d'Informatique de Nantes Atlantique
— Michel Leclère LIRMM (CNRS - UM2)
— Maxime Lefrançois MINES Saint-Etienne
— Alain Leger France Telecom R&D - Research - Orange Labs
— Dominique Lenne Heudiasyc, Université de Technologie de Compiègne
— Cédric Lopez Emvista
— Pascal Molli University of Nantes - LS2N
— Alexandre Monnin Origens Medialab
— Isabelle Mougenot Université Montpellier EspaceDev
— Fleur Mougin ERIAS, INSERM U1219 - Université de Bordeaux
— Amedeo Napoli LORIA Nancy (CNRS - Inria - Université de Lorraine) France
— Jérôme Nobécourt LIMICS

— Nathalie Pernelle LRI-Universit Paris SUD
— Camille Pradel Synapse Développement
— Yannick Prié LINA - University of Nantes
— Cédric Pruski Luxembourg Institute of Science and Technology
— Sylvie Ranwez LGI2P / Ecole des mines d'Alès
— Chantal Reynaud LRI, Univ. Paris-Sud, CNRS, Université Paris-Saclay
— Catherine Roussey Irstea Clermont-Ferrand Center
— Fatiha Saïs LRI (Paris Sud University &CNRS8623), Paris Saclay University
— Pascal Salembier Université de Troyes
— Karim Sehaba LIRIS - Université Lumière Lyon 2
— Konstantin Todorov LIRMM / University of Montpellier
— Raphaël Troncy EURECOM
— Haifa Zargayouna University Paris 13

Le Web Sémantique des Objets : interopérabilité, mise en œuvre et perspectives

Positionnement sur le Web Sémantique des Objets

Nicolas Seydoux[1,2,3], Maxime Lefrançois[4], Lionel Médini[5]

[1] LAAS-CNRS
7 avenue du colonel Roche, F-31400 Toulouse, France
prénom.nom@laas.fr

[2] IRIT Toulouse, France
prénom.nom@irit.fr

[3] UNIV. DE TOULOUSE, INSA, LAAS, F-31400 Toulouse, France
[4] Mines Saint-Etienne, Univ Lyon, Univ Jean Monnet, IOGS, CNRS, UMR 5516, LHC,
Institut Henri Fayol, F - 42023 Saint-Etienne France
prénom.nom@emse.fr

[5] Univ. Lyon, Université Lyon 1 LIRIS, CNRS UMR5205,
F-69622, Villeurbanne, France
prénom.nom@liris.cnrs.fr

Résumé : L'Internet des Objets (IoT en anglais) est un domaine riche en enjeux scientifiques, technologiques et sociétaux. Afin de faire face aux problématiques d'interopérabilité faisant obstacle au développement de l'IoT, les principes et les technologies du Web Sémantique sont intégrées aux réseaux d'objets connectés. Cette intégration est désignée sous le nom de Web Sémantique des Objets (SWoT en anglais). L'émergence de ce domaine repose sur l'adaptation des technologies du Web Sémantique aux contraintes de l'IoT. Ce papier de positionnement vise à décrire les vocabulaires principaux utilisés dans le SWoT, ainsi que l'architecture dans lequel la pile technologique du Web Sémantique se déploie, afin d'ensuite identifier les problématiques émergentes dans le domaine du SWoT. Ces problématiques sont des domaines d'intérêts potentiels pour les futurs travaux de la communauté Ingénierie des Connaissances (IC).

1 Introduction

Les objets connectés sont désormais une réalité du quotidien, et depuis le premier emploi du terme en 1999, les enjeux scientifiques, techniques et sociétaux autour de la notion d'Internet of Things (IoT) n'ont fait que croître (Ashton, 2009).

En particulier, le déploiement de réseaux dits IoT a des enjeux dans des domaines divers, tels que la ville intelligente, la télé-santé, l'agriculture numérique, ou l'industrie du futur. La grande hétérogénéité de ces domaines d'application a mené à l'émergence de "silos", avec l'intégration verticale de l'ensemble de la chaîne de traitement, depuis l'objet collectant la donnée jusqu'au programme l'intégrant dans son processus de collecte et de décision, sans oublier l'interface de l'application finalement consultée par l'utilisateur. Cette fracturation verticale est une source de problèmes d'interopérabilité et limite donc l'émergence d'applications complexes et multi-domaines, ainsi que l'intégration horizontale d'objets ou de services.

Depuis sa création il y a 30 ans, le Web vise à masquer l'hétérogénéité des plateformes matérielles et logicielles, et à permettre l'utilisation d'applications aussi spécifiques et performantes que possibles, en standardisant les langages et protocoles destinés à chaque type d'application. Dans cet esprit, le World Wide Web Consortium (W3C) s'est intéressé aux silos créés par l'IoT et est actuellement en train de finaliser plusieurs standards concernant le Web des Objets (WoT) [1]. Depuis leurs origines (Berners-Lee *et al.*, 2001), les principes et technologies du Semantic Web (SW) ont été dévolus à la communication entre agents complexes, et ont promu l'interopérabilité. L'utilisation de modèles riches et expressifs semble particulièrement indiquée dans le cas des objets connectés, car on se situe dans le cas d'une

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

communication destinée principalement à être interprétée par une machine, et ce n'est qu'une fois transformée qu'elle sera éventuellement transmise à un utilisateur humain.

La convergence entre les domaines de l'IoT, du WoT et celui du SW ont mené à l'apparition d'un nouveau domaine de recherche , le Semantic Web of Things (SWoT). D'abord représenté par les réseaux de capteurs sémantisés (Sheth *et al.*, 2008), le SWoT s'est ensuite diversifié pour aussi intégrer des actionneurs (Wang *et al.*, 2015). Cependant, l'intégration des technologies du SW dans les réseaux d'objets connectés n'est pas triviale. En effet, ce sont des technologies qui sont en général demandeuses en ressources, autant en termes de puissance de calcul pour le raisonnement qu'en bande passante pour le transfert de données ou qu'en mémoire pour le stockage. Cette consommation importante de ressources est en opposition directe à la nature contrainte des réseaux IoT, dans lesquels sont déployés de nombreux objets aux ressources limitées. En effet, afin de permettre la dissémination d'un grand nombre d'objets dans l'environnement, il est nécessaire de miniaturiser ces objets, ainsi que de potentiellement les rendre autonomes en énergie. La combinaison de ces deux contraintes amène à une limitation nécessaire de la puissance de calcul, de l'espace mémoire, et des capacités de communication de l'objet. De plus, d'après des estimations récentes, le nombre d'objets connectés va continuer à croître rapidement, pour passer d'environ 23 milliards en 2018 à une estimation de 75 milliards en 2025. Cette augmentation importante du nombre d'objets s'accompagne nécessairement de l'augmentation du volume de données qu'ils collectent, amenant un problème de passage à l'échelle pour les technologies du SW.

Partant de ce constat, nous questionnons dans ce papier une possible convergence entre les domaines du SWoT et des objets contraints. En particulier, nous nous appuyons sur des travaux existants dans la communauté IC dans ces deux domaines, en termes d'intérêt et de faisabilité, pour mettre en lumière certaines questions qu'une telle convergence fait émerger. Après une présentation des ontologies principales utilisées pour produire des messages compréhensibles par les machines dans le SWoT, une architecture de référence permettant de situer les contributions du Web Sémantique dans l'IoT est détaillée. Les défis représentés par le déploiement du SWoT, ainsi que des contributions y faisant face, sont ensuite décrits, afin d'identifier de futurs domaines d'intérêt pour la communauté.

2 Interopérabilité entre objets

L'un des moteurs poussant au développement du SWoT est le besoin d'interopérabilité de l'IoT, et en particulier d'interopérabilité sémantique (Murdock et al, 2016). En effet, les données et les systèmes qui les collectent sont extrêmement divers, ce qui rend complexe l'intégration de multiples sources de donnée à grande échelle. Afin de décrire le domaine de l'IoT, de nombreuses ontologies ont été proposées, c'est pourquoi nous avons identifié celles ayant un intérêt particulier. Les ontologies que nous avons listées ici sont celles se rapportant directement aux réseaux d'objets connectés, permettant de décrire les dispositifs, la manière dont ils interagissent avec le monde réel, ou la manière dont leurs services peuvent être découverts ou sollicités. Les domaines d'application de l'IoT sont nombreux, et décrire les ontologies qui s'y rapportent sort du périmètre de ce papier.

2.1 Les ontologies standardisées

Dans les domaines technologiques, le premier recours au manque d'interopérabilité est la standardisation. Avec l'émergence du SWoT, les standards incluent des ontologies, de façon à promouvoir l'utilisation de quelques ontologies de référence, plutôt que d'observer un morcellement et la création de nouvelles ontologies avec chaque nouveau projet. Les organismes de standardisation qui contribuent aux ontologies du SWoT sont notamment :

— Le W3C, organisme international de standardisation du Web et du Web sémantique
— L'European Telecommunication Standards Institute (ETSI), organisme européen de standardisation pour les télécommunications
— oneM2M, consortium international rassemblant des organismes de standardisation (dont l'ETSI), des organismes de recherche et des industriels autour d'un standard pour

l'IoT.

Le groupe d'incubation Semantic Sensor Network du W3C a publié en 2011 un rapport analysant les différents modèles conceptuels existants pour décrire les capteurs et leurs observations, ainsi qu'une proposition d'ontologie SSNX (Lefort *et al.*, 2011). L'ontologie SSNX a été massivement réutilisée par d'autres ontologies et jeux de données. Le groupe de travail Spatial Data on the Web Working Group commun aux organismes de standardisation Open Geospatial Consortium (OGC) et W3C ont récemment publié une mise à jour de cette ontologie, nommée **SOSA/SSN** (Haller *et al.*, 2017; Janowicz *et al.*, 2018; Haller *et al.*, 2018). Celle-ci spécifie la sémantique des capteurs et actionneurs, entre autres. Elle permet de décrire notamment les capteurs, les propriétés des choses d'intérêt qu'ils observent, les observations qu'ils font et le résultat de ces observations. De manière analogue, elle permet de décrire les actionneurs, les propriétés des choses d'intérêt sur lesquelles ils peuvent agir, et les actionnements qu'ils font et le résultat de ces actionnements. L'ontologie SOSA/SSN est modulaire. Par exemple :

— SOSA (Sensor, Observation, Sampler, and Actuator) est un module qui comprend seulement quelques termes et peu d'axiomatisation ;
— SSN (Semantic Sensor Network) importe SOSA, introduit d'autres termes, et enrichit l'axiomatisation de SOSA ;
— SSN-System (SSN System Capability module) importe SSN (et SOSA par transitivité), et définit des classes supplémentaires pour modéliser les capacités des systèmes, leurs domaine d'opération ou leurs conditions nominales de déploiement.

Le W3C comprend un autre groupe de travail qui contribue directement au SWoT : le W3C Web of Things working group [2]. Entre autre, ce groupe a pour tâche la conception d'une ontologie, la **Thing Description ontology**, ainsi que son association à un ensemble de métadonnées décrivant les données exposées aux applications par les objets, leurs règles de sécurité, ou les informations de connexion à leurs points d'accès. La spécification de l'ontologie associée au modèle Thing Description est disponible dans (Kaebisch *et al.*, 2019), encore en cours de développement. Une Thing est définie dans (Kovatsch *et al.*, 2019) comme l'abstraction d'une entité (virtuelle ou physique) devait être manipulée par une application IoT, *e.g.* un objet, un service, ou une entité logique telle qu'une pièce ou un bâtiment.

L'ontologie Smart Appliances Reference (**SAREF**) (Daniele *et al.*, 2015), dont le développement est soutenu par l'ETSI, est centrée sur le concept de Device en tant qu'objet physique remplissant une ou plusieurs fonctions. Initialement conçue pour décrire le domaine de la maison intelligente, SAREF a été étendue par des modules spécialisés pour l'énergie, les bâtiments, l'environnement ou l'usine du futur. (Poveda Villalón *et al.*, 2018) a analysé l'alignement possible entre SAREF et SSN, notamment pour l'étendre à la description d'un scénario d'agriculture numérique.

La **oneM2M Base Ontology** [3] est documentée dans l'une des spécifications techniques [4] du standard international oneM2M par le consortium éponyme. La oneM2M Base Ontology est une ontologie de type "cœur de domaine" : elle ne décrit qu'un ensemble réduit de concepts de haut niveau qui vise à être étendu par les implémentations du standard. Ces concepts, par exemple Device, Service ou Variable, font référence à des éléments du standard. Un alignement entre SAREF et la oneM2M base Ontology est aussi distribué par oneM2M [5]. La prochaine version de SAREF devrait également comporter un alignement avec SOSA/SSN.

Il subsiste des questions ouvertes sur la modélisation et l'utilisation de ces ontologies, voici deux points de départ :

— La section (Haller *et al.*, 2017, §7) discute de questions ouvertes de modélisation avec SOSA/SSN. Les différentes options présentées n'ont pas encore fait l'objet d'articles de recherche.

2. W3C WoT Charter - `https://www.w3.org/2016/12/wot-wg-2016.html`
3. `https://git.onem2m.org/MAS/BaseOntology/raw/master/base_ontology.owl`
4. TS-0012 : `http://www.onem2m.org/technical/published-drafts`
5. `https://git.onem2m.org/MAS/BaseOntology`

— L'ontologie TD étant maintenant stabilisée, il reste à ré-étudier l'alignement et l'interopérabilité possible avec SOSA/SSN.

2.2 Autres ontologies importantes du SWoT

Ces ontologies ne sont pas proposées par des organismes de standardisation, mais sont elle-même construites sur des ontologies de référence, issues d'organismes comme le W3C, ou de standards *de facto* dont le statut est acquis par l'usage.

L'ontologie **SEAS** [6] (Lefrançois, 2017) est une ontologie modulaire dont tous les termes sont contenus dans le même espace de noms. SEAS étend SOSA/SSN, et propose un cœur de quatre patrons ontologiques décrivant les systèmes physiques et leurs connexions, les valeurs associées à leurs propriétés, et les processus par lesquels ces associations de valeurs sont faites. Ces patrons sont ensuite instanciés dans chaque module pour un domaine particulier. Dans le contexte d'un projet ETSI Specialist Task Force 556 [7] ces patrons ontologiques sont partiellement incorporés dans SAREF.

L'ontologie **S3N - Smart Semantic Sensor Networks** [8] (Sagar *et al.*, 2018) est une ontologie modulaire qui étend SSN pour décrire des capteurs intelligents, c'est à dire des plateformes hébergeant au moins un capteur, un dispositif communicant, et un micro-controlleur.

IoT-O [9] est une ontologie modulaire pour l'IoT, construite selon un ensemble de bonnes pratiques visant à permettre sa réutilisation (Seydoux *et al.*, 2015, 2016a). Elle est notamment basée sur SSN initialement, et dispose d'une version mise à jour basée sur SOSA. Les différents modules de IoT-O, e.g. Capteurs et Observation ou Actionneurs et Actions, couvrent chacun un domaine de l'IoT, et sont indépendants les uns des autres.

schema.org est une ontologie légère décrivant un large panel de concepts se rapportant aux ressources publiées en ligne (*e.g.* blogs, boutiques en ligne ou contenu multimédia). Ce cœur a été ensuite étendu avec des vocabulaires se rapportant à un domaine spécifique présent sur le Web, comme les références bibliographiques [10], et en particulier l'IoT, avec **iot.schema.org**. Un point de départ pour suivre le développement de cette extension ou contribuer est le groupe collaboratif GitHub `https://github.com/iot-schema-collab/`.

3 Une architecture de référence pour le SWoT

La maturité du domaine du Web Sémantique, et le développement rapide du domaine de l'IoT, amènent l'émergence de problématiques spécifiques au SWoT. Ces problématiques s'articulent autour d'une dichotomie entre les **ressources nécessaires** au déploiement des technologies du Web Sémantique et les **capacités limitées** des nœuds présents dans les réseaux IoT. Ces contraintes sont caractérisées dans la suite de la présente section par l'introduction d'un patron architectural de référence pour le SWoT, utilisé pour situer les problématiques émergentes. Ce patron est ensuite instancié par une architecture orientée avatar permettant d'envisager la construction d'applications WoT et SWoT.

3.1 Que sont le Cloud et le Fog computing ?

Le Cloud computing (Mell & Grance, 2011) est un paradigme dans lequel les traitements sont effectués à distance sur des machines dotées d'importantes ressources : puissance de calcul, mémoire, stockage, bande passante, voire systèmes et applications. L'**accessibilité via le Web** et l'**élasticité** des machines du Cloud en font des relais intéressants pour les applications

6. `https://w3id.org/seas/`
7. `https://portal.etsi.org//STF/STFs/STFHomePages/STF556`
8. `https://w3id.org/s3n/`
9. `http://irit.fr/recherches/MELODI/ontologies/IoT-O`
10. `https://bib.schema.org/`

du SWoT. Dans ce modèle, les données sont collectées par les objets connectés, puis concentrées sur les nœuds Cloud où elles sont stockées, traitées, et intégrées à des applications.

Cependant, les réseaux d'objets connectés contiennent par nature des objets contraints, et très disséminés dans leur environnement. Leurs capacités de communication limitées peuvent être un obstacle à la communication systématique avec des nœuds distants. De plus, la concentration des données dans un serveur Cloud peut introduire un délai important, dépendant du volume de données qui transitent et de la qualité du lien réseau.

Pour faire face à cette fracture entre les infrastructures Cloud et les sources de données de l'IoT, un nouveau paradigme est apparu, le Fog Computing (Bonomi *et al.*, 2012), basé sur l'utilisation des ressources de calcul et de stockage disponibles **en bordure de réseau**. Les nœuds Fog sont donc **massivement distribués**, **hétérogènes**, et disposent de **capacités de calcul limitées**, et se situent entre les objets connectés et les machines du Cloud. Cette définition englobe notamment les passerelles que l'on retrouve dans de nombreux déploiement d'objets connectés, et qui assurent la connexion entre les protocoles de l'IoT et Internet, permettant ainsi de transporter les données des objets jusqu'aux serveurs distants du Cloud.

Le Cloud comme le Fog computing visent à fournir une abstraction des ressources de calcul pour permettre une gestion plus souple des tâches. Cependant, il existe des différences fondamentales entre ces paradigmes. **(I)** Les nœuds du Fog sont situés **à proximité** des objets connectés, fournissant ainsi des capacités de calcul plus proche des sources de données dans le cas de l'IoT. **(II)** Le traitement distribué supporté par le Fog computing a des propriétés de résilience, par l'absence de point critique sur lequel repose l'intégralité du calcul. **(III)** Cloud et Fog offrent des approches différentes pour le passage à l'échelle. Le Cloud s'appuie sur deux axes : l'attribution de machines supplémentaires à la tâche en cours, ou l'attribution de ressources supplémentaires aux machines qui y sont déjà dédiées. Dans le Fog computing, les ressources des nœuds n'étant pas flexibles comme dans le Cloud, seule l'augmentation du nombre de machines est possible. Cependant, cette augmentation se faisant localement par rapport aux sources de données, et s'appuyant sur des machines faisant potentiellement déjà partie du déploiement d'objet, elle s'intègre naturellement au développement massif des grands réseaux d'objets, amenant des ressources complémentaires à celles des serveurs Cloud (Dastjerdi & Buyya, 2016). **(IV)** La nature distribuée des nœuds constituant l'infrastructure Fog rend aussi plus fluide la répartition de la puissance de calcul face à la mobilité des besoins, ce qui est en particulier adapté au domaine de l'IoT (Dastjerdi & Buyya, 2016), où les objets sont disséminés dans un espace géographiquement étalé dans lequel les utilisateurs et certains objets se déplacent.

Le Fog computing n'a pas vocation à remplacer le Cloud computing, mais bien à le complémenter. La capacité de calcul limité des nœuds Fog, leur mobilité, et la localité de leur déploiement ne sont pas adaptés à la prise en charge de cas d'utilisation du Cloud computing. Le rapprochement de ces deux paradigmes est un vecteur intéressant pour fournir des infrastructures soutenant le développement de l'IoT en général, et du SWoT en particulier.

3.2 Structurer une architecture de référence du SWoT

Tous les nœuds de l'IoT ne sont pas équivalents, et ce domaine se distingue par l'hétérogénéité des capacités des machines impliquées, depuis des objets très contraints jusqu'aux serveurs Cloud dotés d'une très grande puissance de calcul. Les principales caractéristiques distinguant les nœuds sont : la puissance de calcul, la mémoire, le stockage, les capacités de communications, et le type de source d'énergie.

Trois classes de nœuds se détachent par l'analyse de la littérature via ces critères :

— Les **noeuds Cloud**, caractérisés par leur puissance de calcul importante, leur large capacité de communication (nombreux protocoles, connectés en permanence), ainsi que leur mémoire et leur capacité de stockage étendue. Cette classe de nœuds inclut les serveurs déployés dans le Cloud, et peut aussi être étendue aux nœuds dont la puissance est importante relativement aux autres machines du même déploiement, pouvant contextuellement s'appliquer à une station de travail classique. De tels nœuds sont présents dans les architectures mises en jeu dans des travaux tels que (Ben-Alaya *et al.*, 2015; Szilagyi & Wira, 2016; Su *et al.*, 2018), avec des appellations variables.

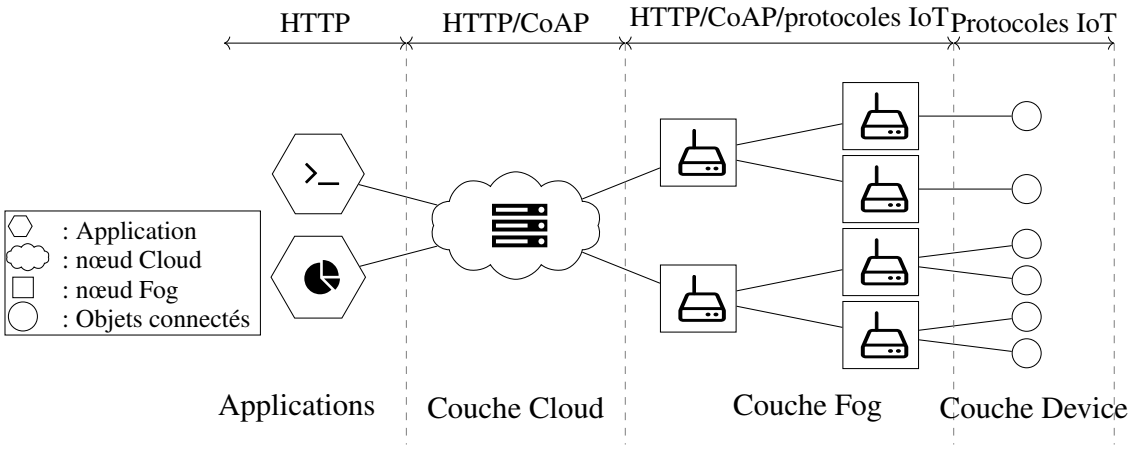


FIGURE 1 – Patron architectural de référence pour le SWoT

— Les **nœuds Fog**, caractérisés en premier lieu par leurs importantes capacités de communication, et leur rôle de pourvoyeur de contenu entre objets connectés et nœuds Cloud en assurant la continuité des protocoles. Ce rôle fait que ces nœuds sont souvent désignés par le terme de "passerelle" dans la littérature (Compton *et al.*, 2009; Ben-Alaya *et al.*, 2015).
— Les **objets contraints**, caractérisés par leurs capacités de traitement et de communication limitée, leur autonomie en énergie, et leur très faible capacité de stockage. Ces nœuds sont par définition présents dans tous les réseaux IoT, dont ils sont la source principale d'informations sur l'environnement.

3.3 Un patron architectural de référence pour le SWoT

Les classes de nœuds précédemment définies constituent un patron architectural récurrent dans la littérature, découpé en trois couches : **Cloud-Fog-Devices**, représenté sur la Fig. 1. Dans ce patron architectural, les services sont exposés aux utilisateurs par un nœud Cloud via des protocoles Web, lequel collecte des données produites par des objets via l'entremise de nœuds Fog. Ce patron architectural, et les différentes couches qu'il définit, seront utilisés comme référence dans la suite de ce papier.

3.4 Composition de services métier orientée-avatars

Depuis les automates programmables et les robots, des éléments de programmation ont été associés aux objets. Avec l'avènement des communications en réseau (sans fil), ils ont évolué vers des systèmes de contrôle distribués, des systèmes embarqués et, plus récemment, l'intelligence ambiante et la robotique distribuée (Galloway & Hancke, 2013). L'IoT est une conséquence directe de cette évolution et vise à exploiter les capacités de communication d'Internet, la puissance de calcul quasi illimitée des infrastructures de la couche Cloud et des interfaces utilisateur modernes pour fournir aux utilisateurs finaux des applications utiles. Le WoT s'appuie sur l'IoT et encourage l'utilisation des standards du Web.

Mais dès que des objets physiques entrent en jeu, la programmation de telles applications devient plus complexe et moins déterministe. Les données des capteurs souffrent d'imprécisions, les actionneurs ont besoin de boucles de rétro-action, les processus critiques et de synchronisation nécessitent une attention constante et les communications réseau peuvent perdre des données ou être interrompues. Les défenseurs des Systèmes Cyber-Physiques (CPS) affirment que de telles difficultés proviennent de phénomènes physiques et doivent être modélisées avec des modèles et des processus informatiques (Lee, 2008). Parmi les nombreuses

architectures CPS qui ont été conçues, bon nombre d'entre elles sont à base de composants et incluent des entités logicielles qui modélisent ces préoccupations.

Dans la communauté WoT, nous avons proposé (Jamont *et al.*, 2014; Mrissa *et al.*, 2015) la notion d'**avatar** pour désigner les artefacts logiciels attachés à un objet et agréger le code nécessaire pour exécuter des applications WoT. Comme détaillé ci-dessous, les avatars reposent sur une architecture interne à base de composants, de sorte que toutes les préoccupations nécessaires du point de vue des CPS puissent être décrites. Plus récemment, le groupe de travail WoT du W3C a proposé la notion de « servient » actuellement définie pour normaliser les objets logiciels dans les applications WoT. Les servients sont très proches des avatars : ils fournissent un accès aux objets, peuvent être exécutés sur ceux-ci dans la couche Device, mais aussi dans les couches Fog et Cloud, et peuvent interagir avec d'autres serveurs. Les deux s'appuient également sur des technologies sémantiques pour échanger des données compréhensibles par une machine. Comme les standards du WoT doivent faire face à une variété de cas d'utilisation et de plateformes, l'architecture cliente ne spécifie que des blocs de construction (« environnement d'exécution », « modèle de ressource », etc.). Les avatars peuvent être considérés comme une spécialisation de servients, davantage centrés sur le déploiement et l'exécution d'applications WoT, et s'appuyant sur une architecture à base de composants pour tirer parti des avancées liées aux préoccupations et aux exigences spécifiques dans divers domaines.

3.5 Avatars

Certains composants de l'architecture de l'avatar (Figure 2) sont dédiés au contrôle des objets et d'autres implémentent le comportement autonome, adaptatif et collaboratif des avatars. La configuration physique est découplée de son architecture logique : un avatar peut adapter de manière dynamique la distribution de ses composants à travers le patron architectural de référence pour améliorer leur efficacité. Nous avons regroupé les composants d'avatar en 8 modules fonctionnels.

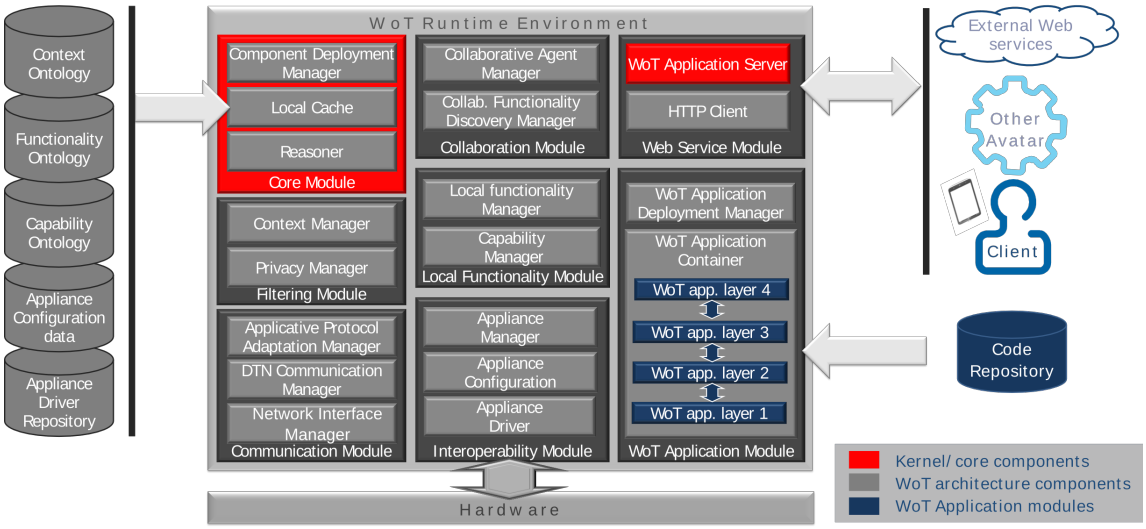


FIGURE 2 – Architecture logicielle de l'avatar

Le **module Core** comprend des composants utilisés à plusieurs étapes du cycle de vie d'un avatar. Le gestionnaire de déploiement de composants définit quels composants d'avatar seront instanciés par rapport aux fonctionnalités de l'objet, et où [11]. Chaque avatar intègre

11. Les composants d'un avatar peuvent être situés : sur l'objet si celui-ci dispose de suffisamment de ca-

un moteur d'inférences, utilisé par d'autres composants pour traiter les informations sémantiques relatives aux capacités, aux fonctionnalités et au contexte. Il en va de même pour le cache local, qui stocke des informations sémantiques provenant de sources diverses (objet, référentiels, contexte externe) et reflète l'état actuel de l'avatar.

Le **module d'interopérabilité** fournit aux autres modules de l'avatar une interface uniforme pour interagir avec l'objet auquel il est attaché. Cette interface consiste en un ensemble de *capacités* représentant l'API de l'objet. Il charge les pilotes depuis un dépôt de la plateforme WoT (voir plus loin) et les utilise pour identifier les schémas de communication compris par l'objet. Finalement, il télécharge sur l'objet la configuration appropriée.

Le **module de filtrage** limite l'exposition des fonctionnalités et les échanges de données. Si, pour des raisons de confidentialité ou de sécurité, certaines fonctionnalités ne sont pas réalisées par l'avatar, elles seront filtrées par le gestionnaire de la confidentialité. Le gestionnaire de contexte a un rôle plus complexe, qui permet notamment de décider quelles fonctionnalités peuvent être exposées, à partir de quelles capacités de l'objet les réaliser, et avec quels autres avatars collaborer pour composer des fonctionnalités complexes (Terdjimi *et al.*, 2016).

Le **module de communication** assure une communication fiable avec l'objet. Il sélectionne l'interface réseau appropriée (Ethernet, Wi-Fi, Zigbee...) et les protocoles (CoAP, HTTP...) en fonction des objectifs de la communication et des besoins en performances (débit / consommation d'énergie). Il prend également en charge les interruptions de connectivité (Mahéo *et al.*, 2012).

Le **module de services Web** permet aux avatars de communiquer entre eux et avec le monde externe dans le respect des standards du Web. Ainsi, les avatars peuvent : interagir avec la plateforme WoT pour interroger des dépôts de code, répondre aux demandes des clients concernant les fonctionnalités qu'ils exposent en tant que ressources RESTful, échanger des données avec d'autres avatars pour réaliser des fonctionnalités collaboratives et interroger des services Web externes pour enrichir leurs propres données.

Le **module de fonctionnalités locales** gère les *fonctionnalités* de haut niveau réalisables à l'aide des capacités atomiques de l'objet et de leur composition.Il s'appuie sur des technologies sémantiques pour lier la couche physique (capacités) avec la couche applicative (fonctionnalités) de manière déclarative et faiblement couplée, en garantissant l'interopérabilité des applications pour les objets (Mrissa *et al.*, 2014). Lorsque l'avatar est créé, le CapabilityManager interroge le module d'interopérabilité pour connaître les fonctionnalités de l'objet et l'ontologie des fonctionnalités de la plateforme pour obtenir leurs descriptions sémantiques. Il est interrogé par le LocalFunctionalityManager, qui charge également les descriptions de fonctionnalités et utilise le raisonneur pour déduire les fonctionnalités locales de l'avatar.Pour chaque fonctionnalité inférée, le LocalFunctionalityManager interroge le gestionnaire de contexte pour déterminer si elle doit être exposée aux clients. Les fonctionnalités exposées sont liées à un registre, afin de permettre leur découverte.

Le **module de Collaboration** gère les fonctionnalités nécessitant une collaboration entre plusieurs avatars. Le CollaborativeFunctionalityDiscoveryManager interroge le raisonneur comme décrit ci-dessus pour identifier, à partir des fonctionnalités locales, celles auxquelles il pourrait participer aux niveaux supérieurs. Ensuite, il interroge le répertoire des fonctionnalités de la plateforme pour rechercher les fonctionnalités manquantes localement. Si de telles fonctionnalités sont disponibles auprès d'autres avatars, il charge le CollaborativeAgentManager, qui gère la négociation avec ces autres avatars, de mettre en oeuvre ces fonctionnalités collaboratives (Jamont, 2016).

Le **module d'Application WoT** fournit et contrôle des « conteneurs d'applications WoT » qui exécutent des modules de code implémentant les différents aspects d'une application (voir ci-dessous). De tels conteneurs peuvent être répliqués sur l'objet, sur la passerelle et sur l'infrastructure cloud grâce au gestionnaire de déploiement, afin que les modules soient exécutés à l'emplacement approprié.

---

pacités de traitement ou pour des modules de code possédant des contraintes temporelles ; dans la couche Fog pour les processus impliquant une communication inter-avatars localisée ; ou dans la couche Cloud pour les processus gourmands en calculs.

3.6 Plateforme orientée-avatar

Les solutions IoT existantes peuvent être utilisées en tant que couche de support pour connecter les objets [12]. Les plateformes WoT sont déployées en tant que « serveurs d'applications WoT » au-dessus de ces solutions. Pour permettre le déploiement et l'exécution d'applications WoT sur des environnements informatiques ubiquitaires, les plateformes WoT doivent fournir un accès à des installations de stockage d'informations et de connaissances et à une puissance de calcul supplémentaire dans la couche Cloud. Dans la mesure où nous promouvons ici les plateformes WoT basées sur des avatars, elles devraient également prendre en charge la gestion, l'exécution et la (dé)sérialisation des avatars. Le mécanisme de sérialisation permet un passage à l'échelle horizontal en répliquant les plateformes et en déplaçant les avatars entre elles à travers les couches Fog et Cloud. La figure 3 décrit l'infrastructure de telles plateformes, où on retrouve instancié le patron architectural décrit en Section 3.3 [13].

Les principaux éléments de cette infrastructure sont le contrôleur d'infrastructure et l'environnement d'exécution WoT. Le premier interagit avec la plateforme IoT et est chargé de décider de créer [14], de mettre à jour [15] ou de supprimer des avatars, au fur et à mesure que les objets sont connectés et déconnectés de la couche IoT. Le second est le conteneur qui isole les avatars et gère leur cycle de vie. Pour des raisons de performances, il est également connecté à la passerelle, de sorte que les avatars puissent directement interagir avec les objets via Fog computing, sans traverser la pile de la plateforme IoT à chaque requête. À l'intérieur du conteneur, les avatars peuvent également : accéder aux ressources Web externes via le proxy Web, partager des informations sur les fonctionnalités exposées par chaque avatar à l'aide du registre de fonctionnalités, interroger les différents dépôts pour accéder aux descriptions sémantiques dont ils ont besoin pour fonctionner, et récupérer le code des modules applicatifs dans le dépôt de code. Les programmes *Device Installer* et *Application Installer* sont chargés d'alimenter les différents référentiels et sont indépendants de la notion d'avatar. Les utilisateurs interagissent en toute sécurité avec la plateforme pour télécharger des pilotes d'objets, installer des applications WoT à partir d'un « marketplace » en ligne ou les exécuter à l'aide de leur navigateur Web, via le proxy qui bloque les demandes entrantes non identifiées.

3.7 Applications WoT

Afin de faciliter la conception des applications WoT et de conserver une indépendance vis-à-vis des caractéristiques des objets disponibles, une application WoT ne traite que la couche fonctionnalité. Elle décrit principalement une hiérarchie de fonctionnalités, dont les nœuds terminaux sont des fonctionnalités terminales (c-à-d. devant être implémentées par une capacité de l'objet [16]), et tous les autres nœuds sont des fonctionnalités composées (*i.e.* qui nécessitent des sous-fonctionnalités et les interrogent à l'aide de modules applicatifs). Certaines de ces fonctionnalités composées peuvent nécessiter d'utiliser les capacités de plusieurs objets et, par conséquent, une collaboration entre plusieurs avatars. La fonctionnalité de niveau supérieur correspond alors à l'application que l'utilisateur final souhaite utiliser.

Une application WoT est packagée dans un fichier compressé, composé de : un fichier *Manifest*, décrivant son contenu ; la hiérarchie de fonctionnalités susmentionnées ; les modules de code correspondant aux algorithmes implémentant les fonctionnalités composées [17] ;

12. Nous avons développé une *Couche d'Interopérabilité Matériels-Applications* (CIMA) sur ce principe, à partir de la plateforme IoT OM2M (Alaya *et al.*, 2014) : `https://github.com/ucbl/CIMA`

13. La Gateway est une passerelle, élément commun de la couche Fog.

14. Chaque avatar est construit de manière à pouvoir accéder à l'objet auquel il se rapporte à travers la passerelle. Les capacités de l'objet sont injectées et les composants de l'avatar instanciés.

15. En requêtant périodiquement la plateforme IoT.

16. Les capacités sont sémantiquement liées à ces fonctionnalités terminales lors de l'initialisation de l'avatar.

17. L'exigence d'interopérabilité impose à ces applications d'être décrites de manière générique, de manière à pouvoir être déployées et exécutées sur différentes configurations. Par conséquent, nous recommandons de

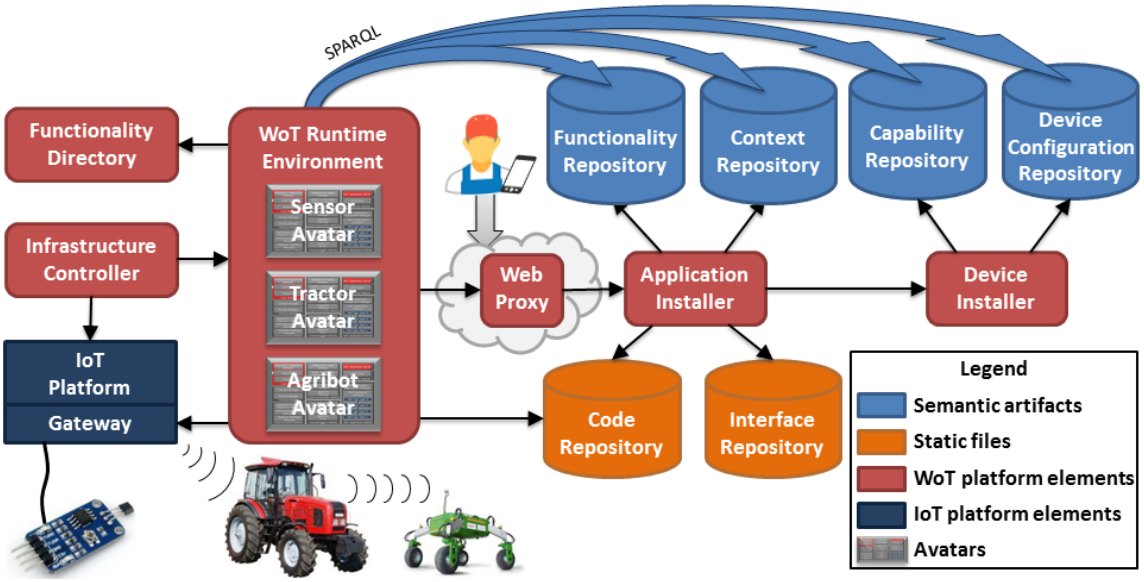


FIGURE 3 – Architecture d'une plateforme orientée-avatar pour le WoT.

le modèle de contexte de l'application, contenant une description sémantique du domaine d'application et un ensemble de règles d'adaptation (Terdjimi *et al.*, 2016) ; un ensemble de fichiers statiques constituant l'interface de l'application et permettant aux utilisateurs finaux d'exécuter et de contrôler l'application via leur navigateur Web en interrogeant les fonctionnalités des avatars à l'aide de standards du Web (ressources RESTful, WebSockets, etc.).

4 Problématiques émergentes

Représenter les données d'une manière qui soit à la fois interopérable et adaptée aux contraintes de l'IoT n'est pas le seul défi auquel fait face le SWoT : l'ensemble des technologies du Web Sémantique est concerné par ce besoin d'adaptabilité. La Section 3 donne l'architecture dans laquelle les composants du SWoT évoluent, ainsi que les fonctionnalités qui y sont regroupées par la notion d'avatar. Cette section montre les défis que représentent l'intégration du SWoT dans cette architecture. En particulier, différent aspects des objets de l'IoT amènent des problématiques dont la communauté du SWoT s'empare peu à peu. C'est notamment le cas de leur nature fortement contrainte, l'aspect éclaté des réseaux qu'ils forment, la forte contextualité des applications qui y sont liées, et les problèmes de passage à l'échelle inhérentes à leur déploiement.

4.1 Combler l'écart entre ontologies et objets

Les ontologies sont des vocabulaires riches, et celles que nous avons identifiées sont toutes représentées en RDF [18]. Basé sur des URI comme identifiants de ressource, le RDF est un langage dont certaines sérialisations peuvent s'avérer assez verbeuses, ce qui implique une consommation d'espace mémoire et de bande passante, qui sont des ressources limitées pour les objets connectés. De ce fait, pour pouvoir manipuler des données sémantisées dans les

les décrire à l'aide de langages déclaratifs, tels qu'une machine à états finis, et lors de l'installation, de les pré-transpiler dans des langages exécutables sur l'objet, sur la passerelle et sur la plateforme cloud sur lesquels ils peuvent être déployés.

18. `https://www.w3.org/RDF/`

couches Fog et Cloud, il est nécessaire de considérer les contraintes de la couche Device. Cette contrainte peut être intégrée aux modules de communication ou de filtrage des avatars. Nous identifions deux approches principales : faire en sorte que les objets puissent directement manipuler du RDF, ou assurer une traduction entre RDF et des formats plus adaptés aux contraintes des objets.

4.1.1 Amener le RDF aux objets...

Afin de rendre possible la manipulation du RDF par les objets, il est nécessaire de proposer des syntaxes RDF plus compactes que les syntaxes les plus répandues, comme le RDF/XML ou la notation N3. (Su *et al.*, 2018) propose un panorama des sérialisations RDF dans l'optique de les manipuler dans un environnement contraint. Par exemple, dans (Charpenay *et al.*, 2018d), les auteurs proposent une sérialisation binaire de JSON-LD afin de permettre l'échange plus efficace des données dans le contexte d'objets contraints. De plus, ce travail est étendu dans (Charpenay *et al.*, 2018b), afin de s'appuyer sur ce format pour stocker et manipuler les bases de connaissance en RDF directement sur les objets connectés. De plus, ces travaux prennent aussi en compte les contraintes en terme de protocoles des communications entre les couches Fog et Device : les protocoles du Web Sémantique, comme HTTP et SPARQL, reposent sur TCP, qui demande l'établissement d'une connexion, ce qui peut être coûteux en bande passante. De la même façon, (Loseto *et al.*, 2016) étend la spécification de Linked Data Platform (LDP) [19], une recommandation du W3C qui décrit un protocole d'exposition de données liées, potentiellement en RDF. Dans la recommandation originale, LDP est associé aux verbes HTTP, que les auteurs ont remplacé par leur équivalent CoAP [20], un protocole construit sur UDP, spécialement conçu pour les échanges dans la couche Device.

4.1.2 ... ou assurer la traduction depuis et vers le RDF

Même si des technologies émergent pour rendre possible la manipulation de RDF sur des objets contraints, un grand nombre d'objets et de services déjà en circulation n'ont pas cette capacité. Pour pouvoir intégrer cette masse d'objets au SWoT, il est donc nécessaire d'établir une transformation entre les formats de données spécifiques à l'IoT et le RDF. Transformer des données vers du RDF est une problématique dont la communauté du Web Sémantique s'est déjà largement emparé : on parle dans ce cas d'**enrichissement**. Cependant, les objets de l'IoT ne sont pas uniquement des producteurs de données, ils en sont aussi des consommateurs, par exemple dans le cas des actionneurs, ou pour la configuration. Il est donc non seulement nécessaire d'enrichir les données issues des objets, mais aussi d'effectuer la transformation inverse, à laquelle on pourrait faire référence comme un **appauvrissement**. En effet, la transformation du RDF vers un format moins expressif s'accompagne d'une perte de contexte, qui n'est plus explicite, mais qui correspond au contexte implicite dans lequel la donnée ainsi appauvrie sera consommée par l'objet destinataire de la communication.

La problématique de la traduction bidirectionnelle entre RDF et formats contraints a été abordée dans divers travaux récents. Dans (Seydoux *et al.*, 2016b), l'approche proposée se destine principalement aux formats arborescents aux schémas explicites, de type XML ou JSON. À partir d'associations entre les balises du langage et les ressources d'un graphe RDF, le système effectue la transformation de manière semi-automatique. Il est particulièrement adapté aux langages formalisés par un standard, dans lequel le schéma est fixé. Les travaux de (Charpenay *et al.*, 2018c) visent eux aussi à assurer la traduction entre différents modèles standards en passant par l'enrichissement. Dans ce cas, la traduction n'est considérée que dans le sens de l'enrichissement. La méthode vise à offrir de l'interopérabilité entre les standards en construisant de manière semi-automatique des règles de traduction d'un modèle enrichi à l'autre. Dans (Lefrançois, 2018), une approche plus souple est proposée en s'ancrant sur les principes de l'architecture du Web : les agents sur le Web, potentiellement contraints, s'échangent des représentations de graphes RDF sous formes de flux d'octets typés par des

19. `https://www.w3.org/TR/ldp/`

20. `http://coap.technology/`

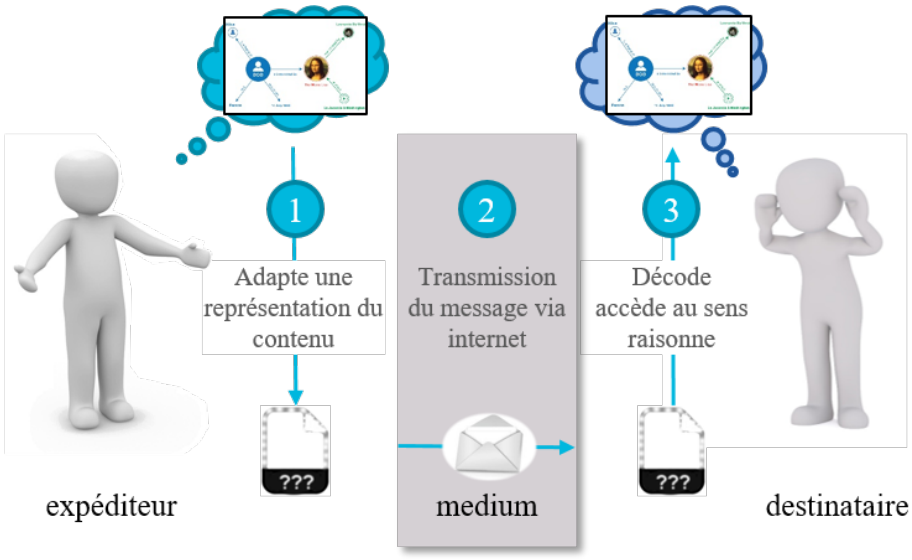


FIGURE 4 – RDF Presentation : utilisation de transformations pour échanger du RDF sur le web (Lefrançois, 2018)

mediatypes, comme l'illustre la Figure 4. Ce cadre conceptuel, nommé RDF Presentation, définit les termes RDF **lowering** (abaissement, étape 1 sur la Figure 4) et **lifting** (élévation, étape 3 sur la Figure 4) au lieu de enrichissement et appauvrissement, étant donné que les représentations RDF sont supposées encoder toutes les caractéristiques du graphe RDF. À partir d'en-têtes HTTP, d'options CoAP, ou d'autres mécanismes, la source ou le destinataire de la communication peut indiquer les règles de transformations entre le format échangé et RDF. Un exemple de langage expressif d'enrichissement ou d'élévation est SPARQL-Generate [21] (Lefrançois *et al.*, 2017), qui est une extension de SPARQL 1.1 qui permet de requêter à la fois du RDF et des documents dans des formats hétérogènes (XML, JSON, CBOR, CSV, HTML, GeoJSON, CBOR, ...) ou des flux de tels documents (WebSocket, MQTT).

4.2 Distribuer le raisonnement

Les technologies du Web Sémantique ont initialement été déployées sur des machines du Cloud pour leur permettre de traiter des données issues d'objets connectés, sans être limitées par leurs contraintes. Le développement du Fog computing offre de nouvelles ressources intégrées aux réseaux d'objets connectés, avec une couche Fog mobile et plus proches de objets. Bien que les nœuds Fog soient largement moins puissants que les serveurs du Cloud, ils peuvent cependant supporter dans une mesure raisonnable les technologies du Web Sémantique sans adaptation particulière de protocole ou de format de donnée. De ce fait, la couche Fog est un candidat privilégié dans le rôle de médiateur entre les formats et les protocoles les plus répandus du Web Sémantique, et ceux spécifiques à la couche Device de l'IoT. La traduction proposée dans la Section 4.1.2 est de ce fait souvent poussée vers les nœuds Fog, qui ont le double avantage de pouvoir communiquer avec les objets connectés, et de gérer un nombre limité d'objet. Il est donc possible pour les passerelles Fog d'utiliser leur connaissance du contexte et des objets auxquels elles sont connectées pour enrichir les données issues des objets, capturant dans la représentation de ces données leur contexte de collecte. Ainsi, les données peuvent être enrichies sémantiquement à partir de leur entrée dans la couche Fog.

Le découpage des avatars en modules de fonctionnalité, et la mobilité de ces modules, permet d'envisager leur distribution à travers l'architecture de référence. En particulier, la des-

21. `https://w3id.org/sparql-generate`

cription des données à partir d'ontologies permet le déploiement de traitements tirant parti du formalisme ainsi introduit, en particulier du raisonnement. Un raisonneur peut, à partir d'une base de connaissance, identifier des inconsistances ou dériver de nouvelles connaissances. La production de nouveaux triplets peut se baser sur des règles d'inférence génériques capturées dans les langages de modélisation comme OWL, mais peut aussi être décrite par des règles d'inférence métier, décrites dans des langages spécifiques. En particulier, SHACL [22] est la recommandation la plus récente du W3C en matière de représentation de contrainte sur des graphes RDF, et une de ses extensions porte sur la représentation de règles [23]. En s'appuyant sur des requêtes SPARQL CONSTRUCT, SHACL permet d'écrire des règles d'inférence menant à la production de nouvelles connaissances. Une fois sérialisées, les règles de déduction peuvent être échangées, rendant possible la distribution du raisonnement, notamment à travers la couche Fog. Dans (Seydoux *et al.*, 2018a,b), les applications consommant des données issues d'un réseau d'objets sont représentées sous la forme d'une collection de règles, lesquelles sont ensuite propagées de proche en proche entre les nœuds du réseau dans une approche nommée EDR. La propagation est complètement décentralisée, et elle est guidée par une définition de politique directement embarquée dans la règle, rendant l'approche complètement flexible. Un exemple de politique visant à amener les règles au plus près des objets est détaillé dans (Seydoux *et al.*, 2018b). Les nœuds échangent entre eux des informations concernant les types de données qu'ils produisent, et les règles sont annotées pour identifier les types de données qu'elles consomment. Les nœuds transmettent la règle aux autres nœuds produisant l'ensemble des données qu'elle consomme, jusqu'à arriver à un point critique, qui est l'ancêtre commun le plus proche des différents producteurs de données nécessaires à la règle. Pour assurer un placement des règles adaptatif aux évolutions du réseau, les mises à jour de la topologie sont propagées dynamiquement par les nœuds, qui remettent en permanence en question la distribution des règles. EDR suppose une disposition hiérarchique des nœuds du réseau en un graphe acyclique, que l'on retrouve dans le patron présenté en Section 3.3. La distribution du traitement permet de réduire la quantité d'information à traiter par chaque nœud appliquant la règle, amenant ainsi à des temps de raisonnement plus court, réduisant ainsi le délai entre l'observation d'un ensemble de signes caractérisant un phénomène et sa prise en compte par une application de haut niveau. De plus, distribuer les règles permet de contextualiser leur application, en ne les considérant que dans les sous-ensemble du réseau dans lequel elles sont pertinentes.

4.3 Distribuer les données

Par nature, les données issues des réseaux d'objets sont fortement liées à un contexte spatio-temporel, exprimant des observations n'ayant de valeur de vérité que si elles sont considérées pour un lieu et un instant donné. Ces données sont produites par des sources très distribuées géographiquement, lesquelles sont potentiellement opérées par des entités indépendantes les unes des autres. De plus, quand bien même les données sont enrichies, soit directement par les objets par l'utilisation des formats spécifiques (*c.f.* Section 4.1.1) soit par l'entremise de nœuds tiers (*c.f.* Section 4.1.2), les ontologies utilisées par les différentes parties ne sont pas nécessairement les mêmes, ou alignées entre elles. Pour permettre l'accès aux données, deux axes d'approche sont possibles : leur duplication, ou leur fédération.

4.3.1 La duplication de données

Dans certains cas, les données issues de sources multiples sont stockées par une entité intermédiaire qui les expose ensuite aux requêtes client. Les données sont donc potentiellement toujours disponibles dans leurs sources respectives, et elles sont dupliquées par l'intermédiaire, ce qui fait que le client interroge un interlocuteur unique pour accéder à des données issues de sources multiples. Ce modèle est instancié par le projet FIESTA-IoT (Sánchez *et al.*, 2018), un projet H2020 visant à exposer via une plateforme unique des données liées issues

22. `https://www.w3.org/TR/shacl/`
23. `https://www.w3.org/TR/shacl-af/`

d'un ensemble hétérogène de sources, telles que des villes intelligentes ou des bâtiments instrumentés. Toutes les sources ne produisant pas leurs données en RDF, n'étant pas nécessairement accessibles en continu, et ne stockant pas nécessairement d'historique, les clients viendront directement interroger la plateforme FIESTA-IoT. Celle-ci se propose donc de collecter les données dans leur format original, de les enrichir à l'aide de leur propre vocabulaire, et de les stocker dans son propre espace mémoire pour fournir un historique aux utilisateurs. FIESTA-IoT est donc hébergée sur un nœud Cloud. Dans le cas de systèmes produisant déjà leurs données en RDF, là aussi la plateforme vient les collecter et les stocker dans son propre espace, à condition que celles-ci soient annotées avec l'ontologie prévue par la plateforme. Le problème de l'interopérabilité est donc résolu par l'utilisation d'un vocabulaire unique au sein de la plateforme. Cependant, les sources de données peuvent tout à fait continuer leur activité de manière indépendante après leur intégration à la plateforme. C'est par exemple le cas du bâtiment ADREAM, au LAAS-CNRS, dont les données sont enrichies pour être publiées sur un portail de données ouvertes [24], en plus d'être publiées dans la plateforme FIESTA annotées avec un vocabulaire différent.

4.3.2 La fédération de données du SWoT

Dans le cas de requêtes fédérées sur un ensemble de sources de données, contrairement au cas précédent, les données ne sont pas dupliquées. C'est le moteur de requête qui va agréger les données issues des différentes sources, afin de fournir une réponse à l'utilisateur. Ce type d'approche ne pose pas le problème de la fraîcheur des données qui peut être observé dans le cas de duplication qui ne soit pas faite en continu, mais par lots de manière ponctuelle. Dans les requêtes fédérées, les sources de données originales sont accédées au moment de la requête. Une telle approche, adaptée au SWoT, est proposée dans (Jaiswal & Lefrançois, 2017). Les auteurs proposent d'étendre la clause SERVICE des requêtes SPARQL, permettant l'interrogation fédérée, afin d'y intégrer les fonctionnalités de SPARQL-Generate. De cette façon, toute source de donnée, même non-RDF, peut être intégrée à une requête fédérée, à la seule condition d'exposer les données à travers un point d'accès HTTP. Cependant, les requêtes fédérées ont aussi leurs limitations, et le choix de l'un ou l'autre des modèles dépend des besoins applicatifs. En particulier, une requête fédérée sera toujours aussi lente que la plus lente des sources de données interrogée, ce qui peut être un problème dans le cas de données directement exposées par des objets parfois peu fiables. De plus, interroger l'ensemble des sources de données à chaque requête peut générer un trafic réseau (et donc une consommation énergétique) plus importante que d'interroger un dépôt centrale dans lequel les données sont stockées après une unique collecte.

4.4 Découvrir les objets et les services

Les réseaux IoT sont des réseaux dynamiques, dans lesquels les changements de topologie dûs à des déplacements, des connexions et des déconnexions de nœuds sont fréquents, notamment dans les couches Fog et Device. Cette accessibilité intermittente des entités amène la question de la découverte : comment, à un instant donné, identifier les nœuds offrant un service utile à une application ? Il est important de remarquer dans ce cas le rôle double des ontologies mises en jeu : non seulement elles décrivent les données métiers manipulées par les objets, mais aussi les objets eux-même. La distribution des données évoquée dans la Section 4.3 peut donc à la fois porter sur les données collectées ou sur la description des objets. Ce second cas est au cœur de la contribution de (Charpenay *et al.*, 2018a). Le cas d'utilisation que présentent les auteurs de ce papier repose sur la description d'objets à partir de SOSA/SSN, et de Thing Description, les deux vocabulaires de référence du W3C. Chaque objet est dépositaire de sa propre description, et pour découvrir l'ensemble des objets et de leurs relations, le client envoie une requête en mode broadcast à l'ensemble du réseau. Il collecte les différentes descriptions envoyées dans les réponses à cette requête, les agrège,

24. `https://syndream.laas.fr`

et exécute l'étape de découverte en attribuant une description sémantique aux nœuds physiques. La question de la découverte des objets peut aussi être comprise dans les protocoles standards. C'est par exemple le cas dans oneM2M, qui définit une procédure de découverte par laquelle un client peut récupérer l'ensemble des ressources connues par un serveur et dont la description est validée par une requête SPARQL.

4.5 Respect de la vie privée

Le développement des réseaux d'objets connectés, et la présence de plus en plus pervasive d'objets dans notre environnement pose d'évidentes questions de respect de la vie privée. Au-delà des questions de sécurisation des objets eux-même, qui est déjà une problématique en soi, les multiples fuites de données récentes amènent à questionner la pertinence de la concentration de masses de données sensibles dans un entrepôt centralisé. C'est pourquoi les approches réparties, très représentées dans les problématiques émergentes (*c.f.* Section 4), sont particulièrement intéressantes. En effet, la distribution du traitement amène un changement de paradigme : les utilisateurs finaux, qui sont ceux qui produisent les données personnelles consommées par leurs fournisseurs de services, peuvent dans ce cas garder la maîtrise de leurs données. Les traitements offerts par les fournisseurs de services sont dans ce type d'approche décorrélés des serveurs contrôlés par ces même fournisseurs, et ne reposent plus sur le transfert de propriété des données par l'utilisateur. En diminuant la concentration des données, le risque de fuite massive est réduit, ainsi que l'attrait pour un potentiel attaquant.

Les ontologies comme acl [25], ou les travaux comme ceux de (Daga *et al.*, 2015), dans lequel les auteurs automatisent la vérification de la conformité des licences d'utilisation des données tout au long de leur processus d'intégration, sont des pistes par lesquelles les principes et les technologies du Web Sémantique peuvent accompagner la distribution du SWoT pour renforcer le respect de la vie privée.

5 Conclusion

L'émergence du SWoT représente un nouveau champ scientifique et technologique où les principes et les technologies du Web Sémantique peuvent représenter un apport majeur. Apporter une réponse aux problèmes d'interopérabilité dûs à l'hétérogénéité des objets et des domaines d'application de l'IoT est le moteur du développement du SWoT. Ce besoin explique l'émergence de plusieurs ontologies et standards visant à faciliter l'interaction entre systèmes. Les déploiements d'objets connectés, de par les contraintes qui leurs sont propres, tendent à reproduire un patron architectural en trois niveaux, Cloud-Fog-Device. Dans ce patron architectural, les composants logiciels permettant le contrôle des objets, que l'on peut structurer sous forme d'avatars, sont distribués selon les besoins et les capacités des nœuds qui les supportent. Le fait que les nœuds composant les réseaux IoT aient des capacités si diverses, et en particulier les objets connectés et les fortes contraintes, est l'aspect qui demande le plus d'adaptation des technologies du Web Sémantique pour développer le SWoT. Pour être manipulables par les objets, le RDF doit être exprimé dans des syntaxes plus compactes et plus légères. Il est aussi nécessaire de prendre en compte l'incapacité de certains objets à manipuler du RDF, et d'assurer dans ce cas une traduction dans les couches Fog ou Cloud. La distribution dynamique du raisonnement et des données est aussi un aspect sur lequel des contributions de la communauté amènent un éclairage nouveau.

L'adoption de vocabulaires et technologies du Web Sémantique dans des standards issus de communautés dans lesquelles ils n'étaient pas d'habitude intégrés est signe de leur démocratisation toujours plus large. Les contraintes de l'IoT, et les innovations qu'elles amènent dans le développement du SWoT, bénéficient au Web Sémantique tout entier, en faisant considérer à la communauté des défis auxquels elle doit faire face, mais aussi en lui donnant accès à une masse de données dont l'exploitation ne peut être que très riche. Les problématiques

25. `www.w3.org/ns/auth/acl`

émergentes identifiées dans ce papier sont autant de domaines d'intérêt pour la communauté IC, et promettent d'intéressants développements dans le futur.

Références

ALAYA M. B., BANOUAR Y., MONTEIL T., CHASSOT C. & DRIRA K. (2014). OM2M : Extensible ETSI-compliant M2M Service Platform with Self-configuration Capability. In *the 5th International Conference on Ambient Systems Networks and Technologies (ANT 2014), the 4th International Conference on Sustainable Energy Information Technology (SEIT-2014)*, p. 1079–1086, Hasselt, Belgium.

ASHTON K. (2009). That 'internet of things' thing. *RFID journal*, **22**(7), 97–114.

BEN-ALAYA M., MEDJIAH S., MONTEIL T. & DRIRA K. (2015). Toward semantic interoperability in oneM2M architecture. *IEEE Communications Magazine*, **53**(12), 35–41.

BERNERS-LEE T., HENDLER J. & LASILLA O. (2001). The Semantic Web. *Scientific American*, **284**(5), 34–43.

BONOMI F., MILITO R., ZHU J. & ADDEPALLI S. (2012). Fog Computing and Its Role in the Internet of Things. In *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*, p. 13–16, New York, New York, USA : ACM Press.

CHARPENAY V., KÄBISCH S. & KOSCH H. (2018a). A framework for semantic discovery on the web of things. In E. DEMIDOVA, A. J. ZAVERI & E. SIMPERL, Eds., *Emerging Topics in Semantic Technologies, ISWC 2018 Satellite Events*, Berlin : AKA Verlag.

CHARPENAY V., KABISCH S. & KOSCH H. (2018b). μRDF Store : Towards Extending the Semantic Web to Embedded Devices. In *ESWC*, p.5̃.

CHARPENAY V., KÄBISCH S. & KOSCH H. (2018c). Semantic data integration on the web of things. In *Proceedings of the 8th International Conference on the Internet of Things, IOT 2018, Santa Barbara, CA, USA, October 15-18, 2018*, p. 3 :1–3 :8.

CHARPENAY V., SEBASTIAN K. & KOSCH H. (2018d). Towards a Binary Object Notation for RDF. In *Extended Semantic Web Conference*, p.1̃5.

COMPTON M., HENSON C., LEFORT L., NEUHAUS H. & SHETH A. (2009). A survey of the semantic specification of sensors. In *International Conference on Semantic Sensor Networks*, volume 522, p. 17–32.

DAGA E., D'AQUIN M., GANGEMI A. & MOTTA E. (2015). Propagation of Policies in Rich Data Flows. In *Knowledge Capture Conference on ZZZ - K-CAP 2015*, p. 1–8.

DANIELE L., DEN HARTOG F. & ROES J. (2015). Created in close interaction with the industry : the smart appliances reference (saref) ontology. In *International Workshop Formal Ontologies Meet Industries*, p. 100–112 : Springer.

DASTJERDI A. V. & BUYYA R. (2016). Fog Computing : Helping the Internet of Things Realize Its Potential. *Computer*, **49**(8), 112–116.

GALLOWAY B. & HANCKE G. P. (2013). Introduction to industrial control networks. *Communications Surveys & Tutorials, IEEE*, **15**(2), 860–880.

HALLER A., JANOWICZ K., COX S. J., LEFRANÇOIS M., TAYLOR K., LE PHUOC D., LIEBERMAN J., GARCÍA-CASTRO R., ATKINSON R. & STADLER C. (2018). Sosa : A lightweight ontology for sensors, observations, samples, and actuators. *Semantic Web Journal*.

HALLER A., JANOWICZ K., COX S. J. D., LE PHUOC D., TAYLOR K. & LEFRANÇOIS M. (2017). *Semantic Sensor Network Ontology*. W3C Recommendation, W3C.

JAISWAL S. & LEFRANÇOIS M. (2017). Towards federated queries for web of things devices. In *Workshop on Semantic Interoperability and Standardization in the IoT, SIS-IoT*, p.4̃p.

JAMONT J., MÉDINI L. & MRISSA M. (2014). A Web-Based Agent-Oriented Approach to Address Heterogeneity in Cooperative Embedded Systems. In J. B. PÉREZ, J. M. C. RODRÍGUEZ, P. MATHIEU, A. CAMPBELL, A. ORTEGA, E. ADAM, E. NAVARRO, S. AHRNDT, M. N. MORENO & V. JULIÁN, Eds., *the 12th International Conference on Practical Applications of Agents and Multi-Agent Systems PAAMS 2014*, volume 293 of *Advances in Intelligent Systems and Computing*, p. 45–52, Salamanca, Spain : Springer.

JAMONT J.-P. (2016). Multi-agent approach, models and tools to collective cyber-physical system engineering. In *Habilitation thesis, Université Grenoble Alpes*.

JANOWICZ K., HALLER A., COX S. J., LE PHUOC D. & LEFRANÇOIS M. (2018). Sosa : A lightweight ontology for sensors, observations, samples, and actuators. *Journal of Web Semantics*.

KAEBISCH S., KAMIYA T., MCCOOL M. & CHARPENAY V. (2019). *Web of Things (WoT) Thing Description*. Candidate Recommendation, W3C.

KOVATSCH, MATTHIASAND MATSUKURA R., LAGALLY M., KAWAGUCHI T., TOUMURA K. & KAJIMOTO K. (2019). *Web of Things (WoT) Architecture*. Candidate Recommendation, W3C.

LEE E. A. (2008). Cyber physical systems : Design challenges. In *Object Oriented Real-Time Distributed Computing (ISORC), 2008 11th IEEE International Symposium on*, p. 363–369 : IEEE.

LEFORT L., HENSON C. & TAYLOR K. (2011). *Semantic Sensor Network XG Final Report*. W3C Incubator Group Report, W3C.

LEFRANÇOIS M. (2017). Planned ETSI SAREF Extensions based on the W3C&OGC SOSA/SSN-compatible SEAS Ontology Patterns. In *Proceedings of Workshop on Semantic Interoperability and Standardization in the IoT, SIS-IoT,*.

LEFRANÇOIS M. (2018). Rdf presentation and correct content conveyance for legacy services and the web of things. In *Proceedings of the 8th International Conference on the Internet of Things*, IOT '18, p. 43 :1–43 :8, New York, NY, USA : ACM.

LEFRANÇOIS M., ZIMMERMANN A. & BAKERALLY N. (2017). A SPARQL Extension for Generating RDF from Heterogeneous Formats. In *ESWC*, p. 35–50 : Springer, Cham.

LOSETO G., IEVA S., GRAMEGNA F., RUTA M., SCIOSCIA F., SCIASCIO E. D. & I B. (2016). Linked Data (in low-resource) Platforms : a mapping for Constrained Application Protocol. In *ISWC*.

MAHÉO Y., LE SOMMER N., LAUNAY P., GUIDEC F. & DRAGONE M. (2012). Beyond Opportunistic Networking Protocols : a Disruption-Tolerant Application Suite for Disconnected MANETs. In *4th Extreme Conference on Communication (ExtremeCom'12)*, p. 1–6, Zürich, Switzerland : ACM.

MELL P. & GRANCE T. (2011). The NIST Definition of Cloud Computing Recommendations of the National Institute of Standards and Technology. *National Institute of Standards and Technology, Information Technology Laboratory*, **145**, 7.

MRISSA M., MÉDINI L., JAMONT J., SOMMER N. L. & LAPLACE J. (2015). An avatar architecture for the web of things. *IEEE Internet Computing*, **19**(2), 30–38.

MRISSA M., MÉDINI L. & JAMONT J.-P. (2014). Semantic Discovery and Invocation of Functionalities for the Web of Things. In *IEEE International Conference on Enabling Technologies : Infrastructure for Collaborative Enterprises*.

MURDOCK ET AL P. (2016). *Semantic Interoperability for the Web of Things*. Rapport interne October.

POVEDA VILLALÓN M., NGUYEN Q.-D., ROUSSEY C., DE VAULX C. & CHANET J.-P. (2018). Ontological requirement specification for smart irrigation systems : A SOSA/SSN and SAREF comparison. In M. LEFRANÇOIS, R. GARCÍA-CASTRO, A. GYRARD & K. TAYLOR, Eds., *Proceedings of the 9th International Semantic Sensor Networks Workshop, International Semantic Web Conference*, volume 2213 of *CEUR Workshop Proceedings*, p. 1–16.

SAGAR S., LEFRANÇOIS M., REBAI I., MAHA K., GARLATTI S., FEKI J. & MÉDINI L. (2018). Modeling smart sensors on top of SOSA/SSN and WoT TD with the semantic smart sensor network (S3N) modular ontology. In E. DEMIDOVA, A. J. ZAVERI & E. SIMPERL, Eds., *Emerging Topics in Semantic Technologies, ISWC 2018 Satellite Events*, Berlin : AKA Verlag.

SEYDOUX N., ALAYA M. B., HERNANDEZ N., MONTEIL T. & HAEMMERLÉ O. (2015). Sémantique et Internet des objets : d'un état de l'art à une ontologie modulaire. In *26es Journées francophones d'Ingénierie des Connaissances*.

SEYDOUX N., DRIRA K., HERNANDEZ N. & MONTEIL T. (2016a). IoT-O, a core-domain IoT ontology to represent connected devices networks. In *EKAW*.

SEYDOUX N., DRIRA K., HERNANDEZ N. & MONTEIL T. (2016b). Lowering knowledge : Making constrained devices semantically interoperable. In *ISWC (Posters and Demos*.

SEYDOUX N., DRIRA K., HERNANDEZ N. & MONTEIL T. (2018a). Towards Cooperative Semantic Computing : a Distributed Reasoning approach for Fog-enabled SWoT. In *COOPIS*.

SEYDOUX N., KHALIL D., HERNANDEZ N. & MONTEIL T. (2018b). A Distributed Scalable Approach for Rule Processing : Computing in the Fog for the SWoT. In *Web intelligence*, Santiago, Chili.

SHETH A., HENSON C. & SAHOO S. S. (2008). Semantic Sensor Web. *IEEE Internet Computing*, **12**(4), 78–83.

SU X., LI P., RIEKKI J., LIU X., KILJANDER J., SOININEN J.-P., PREHOFER C., FLORES H. & LI Y. (2018). Distribution of Semantic Reasoning on the Edge of Internet of Things. In *IEEE UbiComp*, number November, p.̃79.

SZILAGYI I. & WIRA P. (2016). Ontologies and Semantic Web for the Internet of Things - a survey. In *IECON* : IEEE.

IC 2019

SÁNCHEZ L., LANZA J., SANTANA J., AGARWAL R., RAVERDY P., ELSALEH T., FATHY Y., JEONG S., DADOUKIS A., KORAKIS T., KERANIDIS S., O'BRIEN P., HORGAN J., SACCHETTI A., MASTANDREA G., FRAGKIADAKIS A., CHARALAMPIDIS P., SEYDOUX N., ECREPONT C. & ZHAO M. (2018). *Sensors*, **18**(10), 3375.

TERDJIMI M., MÉDINI L., MRISSA M. & LE SOMMER N. (2016). An Avatar-based Adaptation Workflow for the Web of things. In *WETICE 2016*, Paris, France.

WANG F., HU L., ZHOU J. & ZHAO K. (2015). A survey from the perspective of evolutionary process in the internet of things. *International Journal of Distributed Sensor Networks*, **2015**.

Intelligence Artificielle et Santé. Une analyse rétrospective depuis 2010

Jean Charlet[1,2], Sandra Bringay[3,4]

[1] AP-HP/DRCI, Paris, France

[2] Sorbonne Université, INSERM, Univ. Paris 13, LIMICS, Paris, France
Jean.Charlet@upmc.fr

[3] LIRMM UM CNRS, 34080 Montpellier, France
bringay@lirmm.fr

[4] AMIS, UNIVERSITÉ PAUL VALÉRY, Montpellier, France

Résumé : Les Journées francophones d'ingénierie des Connaissances ont hébergé de nombreux articles liés à l'intelligence artificielle et à la santé au sein des sessions plénières mais également dans les ateliers associés. Ces journées ont permis de créer des forums d'échanges scientifiques et de créer une communauté de recherche active. L'objectif de cet article est de réaliser une revue semi-automatique des articles produits dans ce contexte sur la thématique de l'intelligence artificielle et de la santé, d'analyser leur impact sur le domaine et d'identifier les défis à relever. Pour cela, nous avons analysé semi-automatiquement 108 articles publiés entre 2010 et 2018, à partir de leurs titre, mots-clés et résumé. Nous avons identifié sept grands thèmes de recherche dont les principaux sont liés à la construction, l'alignement et l'utilisation d'ontologies, la recherche d'informations, les annotations, les recommandations, les bonnes pratiques, le traitement automatique de la langue, la fouille de données de santé et l'analyse des médias sociaux pour des applications liées à la santé. Cette revue montre l'hybridation des approches symboliques et numériques pour la création de méthodes et d'outils au coeur de la médecine du futur.

Mots-clés : santé, intelligence artificielle

1 Introduction

L'intelligence artificielle est une science, dont le but est de faire faire par des machines des tâches que l'homme réalise grâce à son intelligence. Les chercheurs développent pour cela des approches et techniques multiples, en s'appuyant sur diverses disciplines comme l'informatique, les statistiques et la neuroscience mais également la psychologie cognitive, l'ergonomie, la linguistique, la sociologie, en combinant différents types de traitements de la donnée comme le traitement automatique des langues, la construction d'ontologies, la fouille de données, l'apprentissage automatique...

L'intelligence artificielle est au cœur de la médecine du futur. Ses applications concernent en effet toutes les activités humaines, dont la santé. Elles permettent de répondre à de nombreux défis comme l'amélioration de la qualité des soins, avec les opérations assistées, le suivi des patients à distance, les prothèses intelligentes, les traitements personnalisés, grâce au recoupement d'un nombre croissant de données (big data), l'indexation des connaissances, l'aide à la décision, etc.

Pour répondre à ces défis, la communauté de recherche française s'est structurée avec des évènements annuels. En particulier, les Journées Francophones d'Ingénierie des Connaissances hébergent chaque année de nombreux articles, liés à la santé au sein des sessions plénières mais également dans des ateliers associés. Ces journées ont permis de créer des forums d'échanges scientifiques et de révéler une communauté de recherche active.

Dans cet article, notre objectif est de réaliser une revue semi-automatique des articles produits sur la thématique de l'intelligence artificielle pour la santé, d'analyser leur impact sur le domaine et d'identifier les défis à relever. Pour cela, nous avons sélectionné a) les articles liés à la santé publiés dans la conférence IC depuis 2010 et b) les articles des ateliers « IC et

Santé », « IA et Santé » et « Symposium sur l'Ingénierie de l'Information de santé » (SIIM), qui se se sont succédés tous les ans depuis l'année 2011. Pour commencer l'analyse, nous avons utilisé la méthode probabiliste générative non supervisée Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003). Cette méthode prend en entrée un ensemble de textes et génère automatiquement les thèmes abordés dans ces textes de manière interprétable, sous forme de listes de mots-clés pondérés. LDA a été utilisé pour de nombreuses applications. Dans un contexte similaire, (Wang *et al.*, 2011) et (Wang *et al.*, 2016) ont élaboré un modèle LDA spécialisé utilisant des termes biomédicaux pour explorer la littérature biomédicale. Nous avons identifié six thématiques automatiquement et une dernière manuellement. Ces sept thématiques représentent les centres d'intérêt des chercheurs de la discipline qui hybrident IA symbolique et numérique et ouvrent des perspectives de recherche pertinentes pour améliorer la santé. Notons encore que nous proposons principalement une étude rétrospective et que nous avons fait le choix de ne pas nous projeter dans le futur, principalement pour des raisons de place.

Cet article est organisé de la manière suivante. Dans la section 2, nous décrivons ce qu'est l'IA. Dans la section 3, nous présentons la méthode semi-automatique qui nous a permis de détecter les thèmes ayant émergé dans les publications. Dans la section 4, nous décrivons ces thèmes avant de conclure dans la section 5.

2 Définition de l'IA

L'intelligence artificielle est née dans les années 1950 avec l'objectif de faire produire des tâches humaines par des machines mimant l'activité du cerveau. Face aux déboires des premières heures, deux courants se sont constitués.

Les tenants de l'intelligence artificielle dite forte visent à concevoir une machine capable de raisonner comme l'humain, avec le risque supposé de générer une machine supérieure à l'homme et dotée d'une conscience propre. Cette voie de recherche est toujours explorée aujourd'hui, même si de nombreux chercheurs en IA estiment qu'atteindre un tel objectif est impossible.

D'un autre côté, les tenants de l'intelligence artificielle dite faible mettent en œuvre toutes les technologies disponibles pour concevoir des machines capables d'aider les humains dans leurs tâches. C'est essentiellement cette deuxième approche qui a mobilisé la communauté française travaillant dans le domaine de l'intelligence artificielle et de la santé. Elle a généré de nombreux systèmes spécialisés et performants qui peuplent aujourd'hui notre environnement : créer des profils de futurs « amis » possibles sur les réseaux sociaux, identifier des dates dans les textes pour ordonner des évènements dans des comptes-rendus d'hospitalisation, aider le médecin à prendre des décisions, etc. Ces systèmes ont en commun d'être limités dans leurs capacités d'adaptation : ils doivent être manuellement adaptés pour accomplir d'autres tâches que celles pour lesquelles ils ont été initialement conçus. On distingue deux courants, l'IA symbolique et l'IA numérique qui tendent à s'hybrider ces dernières années (Babbar *et al.*, 2018).

2.1 IA symbolique

L'approche la plus ancienne s'appuie sur l'idée que nous raisonnons en appliquant des règles logiques (déduction, classification, hiérarchisation, etc.). Les systèmes conçus sur ce principe appliquent différentes méthodes, fondées sur l'élaboration de modèles d'interaction entre agents (systèmes multi-agents), de modèles syntaxiques et linguistiques (traitement automatique des langues) ou d'élaboration d'ontologies (représentation des connaissances). Ces modèles sont ensuite utilisés par des systèmes de raisonnement logique pour produire des faits nouveaux.

Dans les années 1980, cette approche, dite symbolique, a permis le développement d'outils capables de reproduire les mécanismes cognitifs d'un expert. C'est pourquoi on les a baptisés « systèmes experts ». Les plus célèbres, Mycin (identification d'infections bactériennes) (Buchanan & Shortliffe, 1985) ou Sphinx (détection d'ictères) (Fieschi *et al.*, 1990), s'appuient

sur l'ensemble des connaissances médicales dans un domaine donné et une formalisation des raisonnements des spécialistes qui lient ces connaissances entre elles pour aboutir à un diagnostic.

Les systèmes actuels, qualifiés d'aide à la décision, de gestion des connaissances ou d'e-santé, sont plus sophistiqués. Ils bénéficient de meilleurs modèles de raisonnement ainsi que de meilleures techniques de description des connaissances médicales, des patients et des actes médicaux. La mécanique algorithmique est globalement la même, mais les langages de description sont plus efficaces et les machines plus puissantes. Ils ne cherchent plus à remplacer le médecin, mais à l'épauler dans un raisonnement fondé sur les connaissances médicales de sa spécialité.

2.2 IA numérique

Contrairement à l'approche symbolique, l'approche dite numérique se focalise sur les régularités qu'il est possible d'identifier dans les données pour en extraire des connaissances, sans modèle préétabli. Cette approche, née avec le connexionnisme (Sun, 1999) et les réseaux de neurones artificiels (Rosenblatt, 1957), se développe aujourd'hui grâce à l'augmentation de la puissance des ordinateurs et à l'accumulation des gigantesques quantités de données, le fameux big data.

La plupart des systèmes actuels procèdent par apprentissage automatique, une méthode visant à donner aux ordinateurs la capacité d'apprendre à partir de données des représentations les plus générales ou spécifiques possibles, qui leur permettent de résoudre des tâches sans être explicitement programmés pour chacune. Par exemple, les algorithmes d'apprentissage profond (deep learning) (Bengio, 2009), dont l'usage explose depuis une dizaine d'années, s'inspirent du fonctionnement cérébral. Ils simulent un réseau de neurones organisés en différentes couches, échangeant les uns avec les autres. La force de cette approche est que l'algorithme apprend la tâche qui lui a été assignée par « essais et erreurs », avant de se débrouiller tout seul. On peut citer par exemple les travaux de (Esteva *et al.*, 2017) qui ont proposé une modèle qui sait reconnaître les mélanomes mieux que 20 dermatologues. La raison s'explique par le nombre de cas (d'images) que le modèle a pu apprendre automatiquement très rapidement, ce qui est impossible pour un humain.

2.3 Hybridation entre IA symbolique et numérique

Ce découpage symbolique/numérique est intéressant car il propose une grille d'analyse des différents travaux que nous voulons étudier. Mais ceux-ci, comme nous le verrons dans la section 4, ne se positionnent pas toujours dans une seule case. Des travaux en IA symbolique utilisent les algorithmes de l'IA numérique pour construire les modèles, par exemple, pour construire des ontologies à partir d'algorithmes numériques de fouille de texte. Des travaux en IA numérique surpassent l'état de l'art en intégrant de la connaissance experte issue de méthodes symboliques, par exemple pour des tâches de classification incluant un raisonnement ontologique (Petegrosso *et al.*, 2016; Notaro *et al.*, 2017). Enfin, cette description n'est pas exhaustive et ne tient pas compte d'un champ spécifique de l'IA, en grande partie absente des travaux que nous analysons, la robotique. Ce champ ferait bien partie de notre analyse mais il est assez spécifique et trop peu d'articles à la base de notre étude s'y intéressent.

3 Analyse semi-automatique du corpus d'articles

L'objectif a été de réaliser un premier tri des articles de manière semi-automatique sous la forme de thèmes que nous avons ensuite interprétés et affinés manuellement.

3.1 Préparation des données

Dans ce travail, nous avons composé un jeu de données de 108 articles. Nous avons sélectionné manuellement 34 articles liés à la santé et publiés dans la conférence IC depuis 2010

ainsi que tous les articles des ateliers « IC et Santé », « IA et Santé » et « Symposium sur l'Ingénierie de l'Information de santé » (SIIM), qui se se sont succédés tous les ans depuis l'année 2011. Cinq ont été exclus car rédigés en langue anglaise. Nous avons récupéré pour chaque article le titre, le résumé et les mots-clés qui ont été concaténés dans une chaîne de caractères. Les mots-clés ont été dédoublés pour leur donner plus de poids et certains ont été normalisés manuellement (e.g. alignement et mapping ont été unifiés en alignement, taln et traitement automatique de la langue en TALN, etc.).

3.2 Application du modèle non supervisée LDA pour détecter les thématiques

Plusieurs modèles permettent d'extraire les thèmes d'un corpus textuel comme l'analyse sémantique latente (Landauer & Dutnais, 1997), l'analyse sémantique latente probabiliste (Hofmann, 2001), l'allocation de Dirichlet latente (LDA) (Blei *et al.*, 2003) et l'indexation sémantique latente (Deerwester *et al.*, 1990). Dans cette étude, nous avons utilisé LDA. C'est un modèle probabiliste avec une définition hiérarchique de ses composantes. Il est génératif, ce qui signifie que nous pourrions générer de nouveaux documents à partir d'un modèle donné. Il est basé sur une représentation relativement simple et robuste des documents textuels. Il ne prend pas en compte l'ordre d'occurrence des termes et la structure des phrases. Le principal avantage de LDA est qu'il s'agit d'un modèle probabiliste avec des sujets interprétables. Son principal inconvénient est qu'il n'existe pas de métrique objective qui justifie le choix des hyperparamètres.

Pour un corpus donné de D documents, nous définissons tout d'abord le vocabulaire pertinent V comme une collection prétraitée de termes occurrant dans le corpus. Dans ce travail, les traitements appliqués sont la mise en minuscules, la suppression des mots vides et des ponctuations, la lemmatisation et la génération des bigrammes et trigrammes. Le corpus obtenu est une collection de D documents $D = \{d_1, ..., d_D\}$. Un document est un N-Uplet de N termes $d = \{t_1, ..., t_N\}$. À chaque terme $t(d, n)$ est associé un thème représenté par la variable $z(d, n)$. θ_d représente la distribution des thèmes du document d. Des hyperparamètres, α et η, définissent l'a priori sur θ et β où β_k décrit la distribution du thème k.

Pour nos expérimentations, nous avons utilisé les librairies NLTK [1] pour les prétraitements ainsi que Gensim [2] et pyLDAvis [3] pour LDA et sa visualisation.

Le nombre de thèmes en entrée de l'approche est difficile à choisir. Dans ce travail, nous avons fait varier le nombre de thèmes K manuellement et retenu 6 thèmes interprétables.

3.3 Visualisation

La principale information que nous pouvons tirer de l'ajustement d'un modèle LDA sur notre corpus de données textuelles est une structure en thèmes ainsi que la répartition des thèmes dans les documents contenus de ce corpus, et l'association de chaque thème à une liste de mots-clés et leur pondération. La figure 1 correspond à la visualisation obtenue en sortie du modèle LDA. Chaque bulle sur le graphique de gauche représente un thème. Plus la bulle est grande, plus ce thème est important. Un bon modèle aura des bulles assez grandes, ne se chevauchant pas, dispersées dans le graphique au lieu d'être regroupées dans un quadrant. Un modèle comportant trop de thèmes comportera des chevauchements et des bulles de petite taille regroupées dans une région du graphique. En sélectionnant une des bulles, les mots et les barres situés à droite sont mises à jour. Ces mots sont les mots-clés dominants du thème sélectionné. Les barres à côté de ces mots-clés correspondent à leur pondération. La visualisation est disponible en ligne [4].

1. https ://www.nltk.org/
2. https ://radimrehurek.com/gensim/
3. https ://github.com/bmabey/pyLDAvis
4. http ://www.univ-montp3.fr/miap/sbringay/lda.html

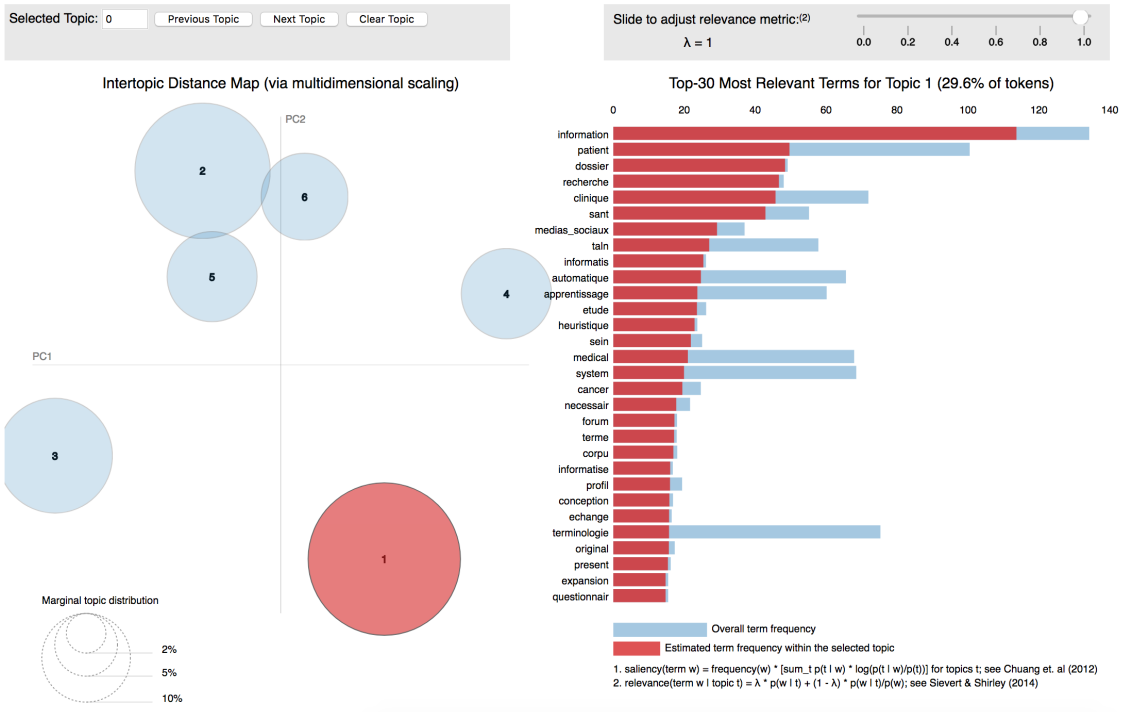


FIGURE *1 – Visualisation des thèmes centrée sur le thème 1 avec les termes associés.*

4 Les sept thèmes principaux liés à l'Intelligence Artificielle et la Santé

L'interprétation des thématiques en sortie du modèle LDA a été réalisée manuellement par les deux auteurs de cet article. Parfois, les mots-clés d'un thème ne sont pas suffisants pour donner un sens à un thème. Nous avons donc consulté les documents associés pour en déduire le sens. Nous décrivons dans la suite les 6 thèmes identifiés par la méthode LDA. Un septième thème a été isolé manuellement par les auteurs comme un sous thème du thèmes 2. Par ailleurs, nous avons réintégré manuellement les articles exclus de l'analyse automatique car rédigés en anglais. Pour finir, quatre articles particuliers (deux très généraux et deux très spécifiques) ont donné lieu à un commentaire dans la section 4.8. Il est important de souligner que cette description n'est pas exhaustive et qu'il existe sûrement d'autres thématiques d'intérêt associant IA et Santé en France, donnant lieu à d'autres publications dans d'autres évènements non étudiés dans cet article.

4.1 Thème 1 : construction d'ontologies et de terminologies

La construction d'ontologies pour la santé est depuis longtemps un des domaines d'application de l'Ingénierie des connaissances, avant même la publication des premiers articles pris en compte dans cette analyse. Les articles de ce thème se focalisent sur le processus de construction de ces ontologies avant la description de l'utilisation de l'ontologie même si celle-ci est généralement renseignée dans les deux cas. Une des spécificités françaises, si ce n'est francophone, est, suite aux travaux du groupe TIA (Terminologie et Intelligence Artificielle) d'avoir mis l'accent sur l'utilisation des corpus textuels pour construire les ontologies.

Ce type de démarche a été utilisé pour les travaux suivants, correspondant chacun à l'élaboration d'une ontologie : une ontologie de domaine en médecine prénatale (Dhombres *et al.*, 2010), une ontologie bilingue de la maladie d'Alzheimer (Dramé *et al.*, 2012), une ontologie de la médecine d'urgence pour l'annotation sémantique (Charlet *et al.*, 2012), une ontologie de la psychiatrie centrée sur l'environnement familial et social du patient (Richard *et al.*,

2013) et enfin une ontologie de la schlérose latérale amyotrophique (SLA – Cardoso *et al.* (2017)). Pas très éloignée de cette démarche, on trouve la construction d'une ontologie des infections orthopédiques avec une approche terminologique à partir des termes de UMLS (De Nizza *et al.*, 2013) et la construction d'une terminologie d'interface avec un processus d'élaboration de libellés précis et valides (Nikiema *et al.*, 2017).

D'autres articles vont mettre davantage l'accent sur les connaissances à l'origine de l'ontologie comme la construction d'une ontologie des maladies infectieuses tenant compte du mode de propagation (Camara *et al.*, 2012) ou l'évolution de l'ontologie des urgences susnommée en tenant compte du modèle d'une ontologie de l'anatomie, le FMA (*Foundational Model of Anatomy* – Charlet *et al.* (2014)). D'autres encore vont insister sur le caractère modulaire de l'ontologie construite, la plupart du temps, à partir d'une ontologie existante déjà décrite auparavant (Camara *et al.*, 2014; Cardoso *et al.*, 2018).

Dans la même veine que la réflexion sur les connaissances, une autre série d'articles va mettre en avant le modèle des données à traiter pour justifier le développement de l'ontologie. Il en est ainsi d'une ontologie dans le domaine de la résistance des bactéries aux antibiotiques (Assele Kama *et al.*, 2010), d'une première ontologie des maladies rares (Dhombres *et al.*, 2011), d'une ontologie pour la qualification et l'indexation des outils d'analyse en sciences omiques (Henry *et al.*, 2016), ou encore d'une terminologie médicale française pour la détection des médicaments en texte libre (Cossin *et al.*, 2018). Dans ce dernier cas, un modèle de données est développé et une ontologie envisagée.

Enfin, à la limite de l'aide à la décision (Cf. thème 4), un dernier article met en avant la sémiologie quantitative des signes modélisés dans le domaine des traumatismes du genou pour justifier la construction de l'ontologie (Guefack *et al.*, 2012).

4.2 Thème 2 : alignement d'ontologies, interopérabilité

Ce thème met en avant des articles liés aux ontologies. Par rapport au thème 1, plus axé sur la construction de l'ontologie, ce thème met l'accent sur l'utilisation de l'ontologie pour l'alignement et l'interopérabilité, le second n'allant pas sans le premier. En effet, l'interopérabilité nécessite d'avoir travaillé en amont sur les questions d'alignement, de façon manuelle ou (semi-)automatisée.

Dans ce contexte, on trouve des articles sur de l'alignement classique, par rapport aux concepts simples : deux travaux visant le même but, à savoir le développement d'alignements multilingues pour le serveur Bioportal du LIRMM (Annane *et al.*, 2016) ou pour le serveur HeTOP (anciennement CISMEF) du CHU de Rouen (Merabti *et al.*, 2012). Avec des méthodes similaires mais fondées sur d'autres terminologies, (Nikiema *et al.*, 2016) vérifient la possibilité de fournir des alignements fiables alors que (Mary *et al.*, 2016) enrichissent les travaux de la même équipe, 4 ans auparavant (Cf. infra), avec de nombreuses méthodes pour fiabiliser au maximum les alignements proposés. Pour des modèles de données plus complexes, (Maaroufi *et al.*, 2014) proposent de formaliser les alignements et de choisir, en fonction de critères formels, quelle technique utiliser. Nous terminons cette revue de détail sur l'alignement par deux articles d'une même équipe sur un problème qui va devenir de plus en plus important au fur et à mesure de l'utilisation des classifications ou ontologies pour l'annotation sémantique : la maintenance de ces annotations malgré l'évolution des ressources utilisées (Reis *et al.*, 2012; Dinh *et al.*, 2013).

Un groupe d'articles s'intéresse aux stratégies mises en place pour assurer le fonctionnement de serveurs de terminologies HeTOP et Bioportal (Jonquet *et al.*, 2010; Grosjean *et al.*, 2011).

Un autre groupe décrit les modèles d'interopérabilité mis en œuvre, avec des représentations plus ou moins liées au web sémantique, pour croiser des données issues des entrepôts de données (Gaignard *et al.*, 2012; Lelong *et al.*, 2017) ou, dans le cadre des maladies rares, la construction de l'ensemble de données minimum de description des patients pour un serveur permettant de fédérer leurs données (Choquet *et al.*, 2012) ou encore, dans la cadre d'une architecture générale pour le diagnostic médical, la proposition d'une méthode d'intégration d'ontologies (Sow *et al.*, 2018).

Pour terminer, (Traore *et al.*, 2014) abordent la question de l'interopérabilité entre dossiers patients hospitaliers à travers des exemples précis tirés d'un projet ANR. (Raboudi *et al.*, 2017) mettent en avant la question de la traçabilité qui est un concept important pour l'interopérabilité dans la mesure ou les données subissent des traitements et des changements qui risquent d'induire des erreurs de transcription et pour lesquels on doit être à même de tracer ces transformations.

4.3 Thème 3 : traitement Automatique de la langue pour le domaine de la santé

Le domaine de la santé est un domaine d'application privilégié pour les méthodes de traitement automatique de la langue naturelle via la création d'outils et de méthodes pour diverses applications visant par exemple l'amélioration des interactions des professionnels de santé avec le dossier patient informatisé via notamment des outils de reconnaissance vocale, l'acquisition de connaissances en matière de santé par les patients, l'identification de patients à partir des textes cliniques, etc.

Un pré-requis à l'analyse des données textuelles des patients est l'anonymisation des identités des personnes et des lieux. (Cardoso *et al.*, 2017) se sont attelés à cette tâche dans le contexte du repérage conceptuel pour la maladie de Charkot. Un autre pré-requis important est la détection de la négation. En effet, les marqueurs de la négation sont constitués d'un ou de plusieurs mots qui modifient la polarité et donc le sens de la phrase (Abdaoui *et al.*, 2017; Dalloux *et al.*, 2017). Pour finir des indices comme la temporalité, le conditionnel ou le possible sont également important à repérer pour l'interprétation (Garcia-Fernandez *et al.*, 2011).

Une fois les pré-traitements effectués, il est alors possible d'extraire de l'information dans les textes comme des co-occurences de concepts médicaux (Abdoune *et al.*, 2011), des signes et des symptômes dans le cas des maladies rares (Martin *et al.*, 2014), des évènements indésirables médicamenteux (Personeni *et al.*, 2016), etc. Ces différents concepts sont évidemment intéressants à relier entre eux (Minard *et al.*, 2011).

Le dossier patient informatisé est alors une source de données textuelles ayant suscité de nombreux travaux supportant l'annotation (Marrast *et al.*, 2013) mais également des raisonnements à partir de ces annotations (Segond *et al.*, 2014) par exemple pour harmoniser la représentation des comptes rendus et évaluer leur similarité (Parès *et al.*, 2014). D'autres types de données, hors dossiers des patients, ont également été utilisés comme les bulletins pour la santé du végétal (Roussey & Ghorfi, 2018) ou encore les requêtes d'experts médicaux (Znaidi *et al.*, 2013).

Parmi les récents challenges à relever, on trouve la prise en compte du multi-linguisme dans les documents relatifs à la santé (Cabot *et al.*, 2017) et la génération de contenus (Kamath *et al.*, 2017) pour répondre à des questions médicales.

4.4 Thème 4 : bonnes pratiques et aide à la décision

Ce thème a été isolé manuellement par les auteurs. En effet, l'analyse du thème 2 a rapidement montré qu'il recouvrait beaucoup plus d'articles que les autres. En approfondissant encore l'exploration, on a vu qu'un certain nombre d'articles correspondaient à l'aide à la décision, thématique historique dans la communauté de l'IA appliquée à la médecine et bien connue des auteurs. Cela nous a donc amené à expliciter ce nouveau thème et nous a permis d'y rattacher 11 articles. On pourrait imaginer qu'une nouvelle paramétrisation des algorithmes de LDA nous aurait permis de retrouver ce thème mais nous n'avions plus de temps pour explorer cette hypothèse.

Un premier groupe d'articles montre le développement de systèmes interactifs d'aide à la décision (SIAD) avec des architectures spécifiques : (Chniti *et al.*, 2012) proposent un SIAD dont les concepts manipulés sont des ontologies avec des règles métiers en JRules ; (Richard *et al.*, 2018) développent une analyse des besoins des médecins avant de proposer un modèle général de raisonnement médical mettant l'accent sur la sélection des informations à donner aux médecins ; (Rybarczyk *et al.*, 2011) développent une plateforme de traitement

multimodal des symptômes aphasiques fondée sur l'analyse des réactions des patients aux divers stimuli.

Un second groupe d'articles propose des SIAD mettant en œuvre des Guides de Bonnes Pratiques (GBP) médicaux : (Meilender *et al.*, 2011) proposent de transformer les recommandations de pratiques cliniques (RPC) en GBP en mettant à disposition un éditeur d'arbre de décision ; (Séroussi *et al.*, 2018) développent un SIAD qui met en œuvre un raisonnement ontologique permettant la réconciliation de différents GBP pour un même patient et présentant les résultats à travers une interface spécifique ; (Schnell *et al.*, 2018) proposent une interprétation des GBP en cancérologie par du raisonnement à partir de cas ; pour finir, (Psiuk & Manet, 2015) analysent la mise en œuvre réelle dans plusieurs hôpitaux du Plan de Soin Type (PST) pour mieux l'implémenter et l'utiliser pour la mise en place de chemins cliniques.

Deux articles analysent des situations de soin pour proposer de nouvelles approches de raisonnement : (Falip *et al.*, 2018) analysent des données médicales de traitement pour proposer un système de recommandation centré sur le patient dans le contexte du raisonnement par analogie ; (Benmouffek *et al.*, 2018) effectuent un travail de recensement et de modélisation de méthodes ayant fait leurs preuves pour l'accompagnement de la prise en charge des personnes avec des troubles neuro-développementaux

Enfin, deux articles abordent des problématiques spécifiques : (Sediri *et al.*, 2012) se fondent sur un modèle de raisonnement pour la gestion de crise pour définir une structure de retour d'expériences et les interfaces permettant de gérer la crise ; (Ugon, 2018) mettent en évidence la difficulté d'évaluer les algorithmes de scorage automatique des données médicales temporelles par rapport aux choix des indicateurs et à la définition des bons *gold standard* auquel se comparer.

4.5 Thème 5 : recherche d'informations, annotations et recommandations

Si le domaine de la recherche d'informations tel que nous le concevons aujourd'hui a émergé dans les années 1950, celui-ci a été rapidement spécialisé dans le domaine de la santé du fait de la nature particulière des documents à indexer et du vocabulaire employé dans ces documents.

Le développement de terminologies et d'ontologies dans le domaine de la santé a permis la mise en place de nouvelles méthodes d'indexation sémantique. La question des modèles a tout d'abord été centrale. Par exemple, (Dinh & Tamine, 2011) ont développé un modèle de recherche d'information multi-terminologique dédié aux documents biomédicaux. (Ghezaiel *et al.*, 2011) ont développé un réseau proxémique pour la recherche d'informations spécifiques à la maladie des dystonies. (Ranwez *et al.*, 2015) ont proposé l'ontologie OntoToxnuc pour représenter la toxicologie nucléaire environnementale.

L'automatisation des annotations sémantiques des textes à partir de ces modèles, est une condition au développement de systèmes de recherche d'information pertinents. Ces annotations sémantiques viennent enrichir les textes et peuvent donc être exploitées ultérieurement comme un point d'accès à la sémantique de ces textes et interrogées sous la forme de divers traitements. (Brut *et al.*, 2011) se sont ainsi intéressés à l'annotation sémantique du dossier patient informatisé. (Dayre & Batatia, 2011) se sont focalisés sur l'annotation de médias selon un modèle de l'activité. (Azzi *et al.*, 2016) ont exploité des annotations sémantiques pour automatiser le calcul des valeurs nutritionnelles d'une recette de cuisine.

Si les textes sont généralement l'objet de la recherche d'information, d'autres médias peuvent représenter la cible de la recherche comme les images dans le système d'information par modalités développé par (Tirilly, 2011) ou encore la recherche sémantique d'images en gastroentérologie (Chabane & Rey, 2013). Une difficulté supplémentaire peut également venir du croisement de différents types de données par exemple cliniques et omiques dans le dossier patient informatisé (Cabot *et al.*, 2015). Pour finir, la disponibilité des données dans le cas de réseaux dégradés (Azanzi *et al.*, 2012) peut également poser la question de l'intégration de différentes sources de données interropérables.

Dernièrement, la recherche d'information a tiré parti des méthodes d'apprentissage pour capturer la sémantique des documents (Nguyen *et al.*, 2017) et raisonner sur ces documents, par exemple pour le codage automatique du PMSI (Ternois *et al.*, 2018), la génération d'alertes

lors de la prescription (Lindemann, 2018), la description de cohortes guidée par les données (Neuraz *et al.*, 2018) ou encore le *rescreening* à partir d'un entrepôt de données cliniques (Pasco *et al.*, 2018).

4.6 Thème 6 : fouille des données de santé

Les méthodes classiques de fouille de données ont été utilisées pour extraire de la connaissance nouvelle à partir de grandes quantités de données de santé. Cette connaissance peut alors être utilisée pour enrichir les interprétations des professionnels de la santé, tout en fournissant des entrées pour des méthodes automatiques ou semi-automatiques exploitant cette connaissance.

On trouve tout d'abord les approches de type recherche de motifs qui vont repérer des régularités généralement fréquentes dans les données : par exemple, (Béchet *et al.*, 2012) utilisent des motifs séquentiels pour découvrir des relations entre des gènes et des maladies rares dans la littérature biomédicale ; (Pinaire *et al.*, 2015b) utilisent ces motifs pour représenter des parcours de soins à partir de données du PMSI ; (Personeni *et al.*, 2018) découvrent des associations entre évènements indésirables médicamenteux.

Si ces motifs sont intéressants pour représenter des données ordonnées (des évènements dans le temps ou des mots dans des données séquentielles textuelles), ils peuvent être étendus pour intégrer de nouvelles dimensions comme, par exemple, la dimension spatiale dans le cas des trajectoires (Pinaire *et al.*, 2017b). Ces motifs ont été utilisés afin de prédire l'évolution des populations de patients et les coûts associés et ainsi recommander les soins les plus efficaces et les moins coûteux (Pinaire *et al.*, 2017a).

D'autres auteurs ont utilisé l'analyse formelle de concepts pour étudier des concepts médicaux lorsqu'ils sont décrits formellement, par exemple pour construire des profils de patients (Séroussi *et al.*, 2013).

D'autres types de données ont également été utilisés comme des images (Abbal *et al.*, 2011) en imagerie ultrasonore ou encore des données textuelles pour parcourir la littérature afin de produire un état de l'art semi-automatique (Thebault *et al.*, 2010; Pinaire *et al.*, 2016).

Les régularités extraites à partir des gros volumes de données médicales vont alors pouvoir être utilisées pour des tâches de prédiction par exemple pour l'aide au diagnostic de tumeurs cérébrales (Ben Lamine *et al.*, 2012), la prédiction des stades de sommeil (Ugon *et al.*, 2016), la réadmission à l'hôpital (Bussy *et al.*, 2018), le diagnostic du diabète (Dendani & Allouani, 2018) ou encore la biométrie de la tête du fœtus (Grandjean *et al.*, 2018).

La visualisation des résultats des méthodes de fouille de données est particulièrement importante pour une bonne appropriation des outils et une interprétabilité des approches par les professionnels de santé. Par exemple, (Pinaire *et al.*, 2015a) ont proposé une visualisation des trajectoires de soins des patients, (Lamy *et al.*, 2017, 2015) une plateforme pour comparer les caractéristiques de médicaments, (Serres *et al.*, 2011) un outil pour le suivi des dissections ou encore pour la visualisation de données multi-sources de chimiothérapie (Jannot *et al.*, 2018).

4.7 Thème 7 : Analyse des médias sociaux

L'extraction de connaissances à partir des données hétérogènes issues des médias sociaux de santé pour des applications liées à la santé est un des thèmes ayant animé la communauté ces dernières années. Ces travaux permettent de prendre en compte de manière originale, les données produites directement par les patients par opposition à des approches où seules les données médicales produites par les professionnels de santé sont étudiées.

Un cadre fédérateur de représentation de ces données, de leur contexte et des indices que l'on peut retrouver dans les médias sociaux a été proposé par (Bricon-Souf *et al.*, 2015; Chahbandarian *et al.*, 2014).

D'autres approches s'intéressent aux thématiques d'intérêt des patients avec des approches d'apprentissage non supervisées sans *a priori* (Pertin *et al.*, 2017) et des approches supervisées nécessitant au préalable la définition de classe d'intérêts (Opitz *et al.*, 2014). L'analyse de ces contenus permet d'identifier des textes liés au risque (Mercadier *et al.*, 2018), à l'incertitude (Abdaoui *et al.*, 2014), aux sentiments (Bringay *et al.*, 2014), à la détresse psycho-

logique (Kessler *et al.*, 2018) ou aux comportements suicidaires (Combes *et al.*, 2016), etc. Ces informations permettent aux professionnels de santé de mieux cerner les perceptions que les patients ont de leur maladie.

Un point récurrent dans ces travaux est la difficulté à mettre en œuvre les chaînes de traitements classiques sur les textes rédigés par les patients qui sont par nature hétérogènes, incertains, entachés d'erreurs et qui nécessitent par conséquent de nombreux pré-traitements. (Tapi Nzali *et al.*, 2015) ont pour cela construit une nouvelle ressource mettant en relation le vocabulaire des patients et celui des professionnels de santé. Cette ressource a été intégrée dans le portail Bioportal [5] qui permet d'annoter des textes de patients avec des concepts médicaux (Eholié *et al.*, 2016).

Pour finir, les aspects temporels ont également été étudiés pour capturer l'évolution de comportements ou de thématiques au cours du temps selon l'histoire des patients. Ainsi, (Moulahi *et al.*, 2017) ont pu mettre en avant des sujets d'interrogation des patients associés à des états d'esprit et à un événement de leur histoire médicale, jusqu'alors méconnus des professionnels de santé dans le cas du suivi et de la détection des idéations suicidaires.

4.8 Articles particuliers

Reste en dehors des thèmes identifiés (6+1), quatre articles, de caractères plus généraux ou avec des spécificités précises. Dès l'instant qu'il apparaissait des articles difficiles à classer, nous n'avons pas cherché à faire disparaître ce « 8^e » thème. Dedans, se trouvent deux articles généraux, celui sur la définition de l'IA par un des auteurs de cet article (Charlet, 2018) et un article qui interroge la gestion des connaissances médicales par rapport à des expériences de terrain (Blanchet *et al.*, 2015). Les deux derniers articles abordent des thématiques très spécifiques : le premier s'intéresse à l'efficacité de la traduction pour améliorer l'accès à l'information médicale dans un contexte transfrontalier (Laforest *et al.*, 2011) et le second teste l'efficacité de routines de déidentification et d'exportation d'images (Seymour & Payoux, 2017).

5 Conclusion, Discussions et Perspectives

Dans cet article, nous avons présenté une étude préliminaire sur les interactions entre le domaine de l'intelligence artificielle et celui de la santé en France. Pour réaliser ce travail, nous avons utilisé une méthode classique de science des données, LDA, pour extraire les principaux thèmes d'intérêts des publications depuis 2010 dans la conférence Ingénierie des connaissances et dans les ateliers associés à cette conférence. Nous avons identifié sept principaux thèmes qui montrent la variété des travaux réalisés ainsi que l'hybridation entre les approches symboliques et numériques. Nous avons aussi fait le choix de ranger chaque article dans un thème unique. Certains d'entre eux pourraient être associés à plusieurs thèmes, en plus ou à la place de celui choisi mais notre but était de mettre en avant les thématiques en prenant chaque article comme exemple illustratif d'une thématique unique.

La principale limitation de cette étude est que nous avons utilisé LDA uniquement comme une entrée à l'interprétation manuelle. LDA nécessite beaucoup de réglages manuels des paramètres. Nous avons passé beaucoup de temps à les identifier pour que les résultats puissent être interprétés de manière significative. En effet, il n'existe pas de métrique objective qui justifie le choix de ces paramètres et en particulier le nombre de thèmes K utilisés en entrée de la méthode, ce qui rend très difficile la généralisation de l'approche à d'autres données et tâches. Par exemple, le thème sur l'aide à la décision a été créé par les auteurs pour reclasser un certain nombre d'articles *a posteriori* du découpage en thèmes fourni par LDA. Par ailleurs, les articles écrits en anglais ont été reclassés à la main. Il est évident également que certains thèmes se recoupent, par exemple le TAL est utile pour la construction d'ontologies ou encore leur alignement. Enfin, cette étude se limite à une seule source d'articles alors que d'autres articles en français liés à l'Intelligence Artificielle et la santé seraient pertinents à

5. http ://bioportal.lirmm.fr/

analyser. Une autre perspective consiste à reproduire l'analyse en travaillant sur des résumés en langue anglaise obtenus via l'API pubmed [6] à partir des mots-clés traduits repérés dans les articles français afin de voir si les tendances se confirment au niveau international. Dans ce contexte, le *YearBook of Medical Informatics* qui parait tous les ans serait une excellence source de textes à analyser puisqu'il est découpé en chapitres précis correspondant pour la plupart à des thématiques de l'IA, comme par exemple (Dhombres & Charlet, 2018). Pour finir, il semble également important de prendre en compte des aspects temporels pour établir une chronologie de l'évolution des thèmes.

Références

ABBAL R., BASARAB A. & KOUAMÉ D. (2011). Estimation de décalages subpixéliques en imagerie ultrasonore. In L. TAMINE, S. DARMONI & L. SOUALMIA, Eds., *Actes de SIIM 2011, 1er Symposium sur l'Ingénierie de l'Information Médicale*, p. 111–118, Toulouse. http ://www.irit.fr/SIIM/SIIM2011-10.pdf.

ABDAOUI A., AZÉ J., BRINGAY S., PONCELET P. & GRABAR N. (2014). Analyse des messages des patients et des médecins dans les fora de santé. In S. BRINGAY, N. SOUF & L. TAMINE-LECHANI, Eds., *Actes de l'Atelier IC Santé*, Clermont-Ferrand. https ://www.lirmm.fr/ic-sante/pmwiki/docs/Abdaoui.pdf.

ABDAOUI A., TCHECHMEDJIEV A., DIGAN W., BRINGAY S. & JONQUET C. (2017). French ConText : Détecter la négation, la temporalité et le sujet dans les textes cliniques Français. In L. TAMINE-LECHANI, Ed., *Actes de SIIM 2017, 4e Symposium sur l'Ingénierie de l'Information Médicale*, p. 7–16, Toulouse. https ://www.irit.fr/SIIM/2017/SIIM2017$_{p}aper_{7}.pdf$.

ABDOUNE H., SOUALMIA L. & JOUBERT M. (2011). Analyse de cooccurrences de concepts biomédicaux dans. In L. TAMINE, S. DARMONI & L. SOUALMIA, Eds., *Actes de SIIM 2011, 1er Symposium sur l'Ingénierie de l'Information Médicale*, p. 71–76, Toulouse. http ://www.irit.fr/SIIM/SIIM2011-6.pdf.

ANNANE A., EMONET V., AZOUAOU F. & JONQUET C. (2016). Réconciliation d'alignements multilingues dans BioPortal. In N. PERNELLE, Ed., *27èmes Journées Francophones d'Ingénierie des Connaissances - IC 2016*, Montpellier, France.

ASSELE KAMA A., MELS G., CHOQUET R., CHARLET J. & JAULENT M.-C. (2010). Une approche ontologique pour l'exploitation de données cliniques. In S. DESPRÈS, Ed., *21èmes Journées Francophones d'Ingénierie des Connaissances - IC 2010*, p. 183–194, Nîmes, France : Ecole des Mines d'Alès.

AZANZI F. J., LO M. & TCHUENTE M. (2012). Vers une approche d'intégration de données adaptée aux réseaux dégradés : application au système d'information sanitaire camerounais. In M.-C. JAULENT & L. SOUALMIA, Eds., *Actes de l'Atelier IC pour l'Interopérabilité Sémantique dans les applications en e-Santé*, Paris. https ://sites.google.com/site/icisante2012/programme.

AZZI R., SYLVIE D. & NOBECOURT J. (2016). Approche sémantique pour automatiser le calcul des valeurs nutritionnelles d'une recette de cuisine. In F. MOUGIN, A. ABDAOUI & P. ZWEIGENBAUM, Eds., *Actes du 2e Atelier IA & Santé*, Montpellier. https ://ic2016.sciencesconf.org/conference/ic2016/pages/IA_Sante.pdf.

BABBAR P., SINGHAL A., YADAV K. & SHARMA V. (2018). Connectionist model in artificial intelligence. *International Journal of Applied Engineering Research*, **13**(7), 5154–5159.

BÉCHET N., CELLIER P., CHARNOIS T. & CRÉMILLEUX B. (2012). Fouille de motifs séquentiels pour la découverte de relations entre gènes et maladies rares. In S. SZULMAN, Ed., *23èmes Journées Francophones d'Ingénierie des Connaissances - IC 2012*, p. 149–164, Paris, France.

BEN LAMINE F. F., KALTI K. & MAHJOUB M. A. M. (2012). Etude de Modèles à base de réseaux Bayésiens pour l'aide au diagnostic de tumeurs cérébrales. In S. SZULMAN, Ed., *23èmes Journées Francophones d'Ingénierie des Connaissances - IC 2012*, Paris, France.

BENGIO Y. (2009). Learning deep architectures for ai. *Found. Trends Mach. Learn.*, **2**(1), 1–127.

BENMOUFFEK D., PACINI L., HONION J., REYDON H. & KERBIRIOU C. (2018). Données de prise en charge pluridisciplinaire des personnes avec TND. In J. CHARLET, M.-D. DEVIGNES & B. SEROUSSI, Eds., *Actes du 3e Atelier IA & Santé*, p. 75–80, Nancy. http ://pfia2018.loria.fr/communicationsiasante_3juillet2018/.

BLANCHET P., REUN R. L. & MORVAN T. (2015). Partage de connaissances médicales par les systèmes automatisés : les besoins des acteurs de terrain. In N. BRICON-SOUF, Ed., *Actes de SIIM 2015, 3e Symposium sur l'Ingénierie de l'Information Médicale*, p. 32–42, Rennes. https ://www.irit.fr/SIIM/2015/ActesSIIM2015.pdf.

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, **3**, 993–1022.

BOUAUD J., SAUQUET D., GIRAL P., JULIEN J., CORNET P., FALCOFF H. & SÉROUSSI B. (2010). Pourquoi les médecins ne suivent-ils pas les systèmes de recommandations de bonnes pratiques ? Une hypothèse liée à l'utilisabilité évaluée avec le mode guidé d'ASTI.

BOUAUD J. & SÉROUSSI B. (2012). Médecine factuelle et recommandations de bonne pratique : une extension du modèle classique pour expliquer les décisions médicales non conformes. In S. SZULMAN, Ed., *23èmes Journées Francophones d'Ingénierie des Connaissances - IC 2012*, p. 251–266, Paris, France.

BOUAUD J., SOULET A., SPANO J.-P., LEFRANC J.-P., COJEAN-ZELEK I., BLASZKA-JAULERRY B., ZELEK L., DURIEUX A., TOURNIGAND C., MESSAI N., ROUSSEAU A. & SÉROUSSI B. (2014). Quels sont les patients atteints d'un cancer du sein dont la décision de prise en charge thérapeutique bénéficie de

6. https ://www.ncbi.nlm.nih.gov/pubmed/

l'utilisation d'un système d'aide à la décision ? Un exemple utilisant la fouille de données et OncoDoc2. In C. FARON-ZUCKER, Ed., *25èmes Journées Francophones d'Ingénierie des Connaissances - IC 2014*, p. 107–118, Clermont-Ferrand, France. Session 2 : Utilisateurs et usages.

BRICON-SOUF N., CHANBANDARIAN G. & HO-DAC M. (2015). Un cadre fédérateur de représentation des données et indices issus de forums de santé. In N. BRICON-SOUF, Ed., *Actes de SIIM 2015, 3e Symposium sur l'Ingénierie de l'Information Médicale*, p. 43–48, Rennes. https ://www.irit.fr/SIIM/2015/ActesSIIM2015.pdf.

BRINGAY S., KERGOSIEN E., POMPIDOR P. & PONCELET P. (2014). Identifier la cible des émotions dans les forums de santé. In C. FARON-ZUCKER, Ed., *25èmes Journées Francophones d'Ingénierie des Connaissances - IC 2014*, p. 163–174, Clermont-Ferrand, France. Session 3 : Web social.

BRUT M., AL-KUKHUN D., PÉNINOU A., CANUT M.-F. & SÈDES F. (2011). Structuration et Accès au Dossier Médical Personnel : approche par ontologies et politiques d'accès XACML. In L. TAMINE, S. DARMONI & L. SOUALMIA, Eds., *Actes de SIIM 2011, 1er Symposium sur l'Ingénierie de l'Information Médicale*, p. 77–86, Toulouse. http ://www.irit.fr/SIIM/SIIM2011-7.pdf.

B. G. BUCHANAN & E. H. SHORTLIFFE, Eds. (1985). *Rule-Based Expert Systems : The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA : Addison-Wesley.

BUSSY S., VEIL R., LOOTEN V., BURGUN A., GAÏFFAS S., GUILLOUX A., RANQUE B. & JANNOT A.-S. (2018). Design d'un algorithme d'IA en grande dimension pour prédire la réadmission à l'hôpital. In J. CHARLET, M.-D. DEVIGNES & B. SEROUSSI, Eds., *Actes du 3e Atelier IA & Santé*, p. 1–7, Nancy. http ://pfia2018.loria.fr/communicationsiasante_3juillet2018/.

CABOT C., SOUALMIA L. F. & DARMONI S. J. (2015). Intégration de données cliniques et omiques pour la recherche d'information dans le Dossier Patient Informatisé. In M.-H. ABEL, Ed., *26èmes Journées Francophones d'Ingénierie des Connaissances - IC 2015*, Rennes, France.

CABOT C., SOUALMIA L. F. & DARMONI S. J. (2017). CIM-IND : Un système multilingue pour l'extraction d'information dans les textes cliniques. In L. TAMINE-LECHANI, Ed., *Actes de SIIM 2017, 4e Symposium sur l'Ingénierie de l'Information Médicale*, p. 105–112, Toulouse. https ://www.irit.fr/SIIM/2017/SIIM2017$_paper_1.pdf$.

CAMARA G., DESPRÉS S., DJEDIDI R. & LO M. (2012). Vers une ontologie des processus de propagation des maladies infectieuses. In S. SZULMAN, Ed., *23èmes Journées Francophones d'Ingénierie des Connaissances - IC 2012*, p. 99–114, Paris, France.

CAMARA G., DESPRÈS S. & LO M. (2014). IDOSCHISTO : une extension de l'ontologie noyau des maladies infectieuses (IDO-Core) pour la schistosomiase. In C. FARON-ZUCKER, Ed., *25èmes Journées Francophones d'Ingénierie des Connaissances - IC 2014*, p. 39–50, Clermont-Ferrand, France. Session 1 : Construction, peuplement et exploitation d'ontologies.

CARDOSO S., AIME X., MEININGER V., GRABLI D., COHEN K. B. & CHARLET J. (2018). De l'intérêt des ontologies modulaires. Application à la modélisation de la prise en charge de la SLA. In S. RANWEZ, Ed., *29èmes Journées Francophones d'Ingénierie des Connaissances - IC 2018*, p. 121–128, Nancy, France : AFIA.

CARDOSO S., AIMÉ X., MORA L. F. M., GRABLI D., MEININGER V. & CHARLET J. (2016). Les ontologies pour aider à comprendre les parcours de santé dans le cadre des maladies neurodégénératives. In F. MOUGIN, A. ABDAOUI & P. ZWEIGENBAUM, Eds., *Actes du 2e Atelier IA & Santé*, Montpellier. https ://ic2016.sciencesconf.org/conference/ic2016/pages/IA_Sante.pdf.

CARDOSO S., MELO MORA L. F., JAULENT M.-C., AIMÉ X., GRABLI D., MEININGER V. & CHARLET J. (2017). Ontologie et TALN : l'anonymisation au service du repérage conceptuel dans le contexte de la SLA. In C. ROUSSEY, Ed., *28èmes Journées Francophones d'Ingénierie des Connaissances - IC 2017*, p. 98–103, Caen, France.

CHABANE Y. & REY C. (2013). Annotation et recherche semantique d'images en gastroenterologie. In L. TAMINE-LECHANI & L. SOUALMIA, Eds., *Actes de SIIM 2013, 2e Symposium sur l'Ingénierie de l'Information Médicale*, Lille. https ://www.irit.fr/SIIM/2013/03_siim13.pdf.

CHAHBANDARIAN G., BRICON-SOUF N., BASTIDE R. & STEINBACH J.-C. (2017). Stable Feature Selection Approach : Application to the Encoding of Secondary Diagnoses. In L. TAMINE-LECHANI, Ed., *Actes de SIIM 2017, 4e Symposium sur l'Ingénierie de l'Information Médicale*, p. 77–84, Toulouse. https ://www.irit.fr/SIIM/2017/SIIM2017$_paper_1 1.pdf$.

CHAHBANDARIAN G., MOJAHID M. & BRICON-SOUF N. (2014). Contextual presentation of medical forum's discussions. In S. BRINGAY, N. SOUF & L. TAMINE-LECHANI, Eds., *Actes de l'Atelier IC Santé*, Clermont-Ferrand. https ://www.lirmm.fr/ic-sante/pmwiki/docs/Chahbandarian.pdf.

CHARLET J. (2018). IA en santé. Dénitions, réalisations et perspectives. In J. CHARLET, M.-D. DEVIGNES & B. SEROUSSI, Eds., *Actes du 3e Atelier IA & Santé*, p. 9–15, Nancy. http ://pfia2018.loria.fr/communicationsiasante_3juillet2018/.

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

CHARLET J., MAZUEL L., DECLERCK G., MIROUX P. & GAYET P. (2014). Décrire les maladies localisées et la physiopathologie pour une ontologie des urgences : un algorithme générique à partir de la FMA. In C. FARON-ZUCKER, Ed., *25èmes Journées Francophones d'Ingénierie des Connaissances - IC 2014*, p. 15–26, Clermont-Ferrand, France. Session 1 : Construction, peuplement et exploitation d'ontologies.

CHNITI A., BOUSSADI A., ALBERT P. & CHARLET J. (2012). Validation pharmaceutique des prescriptions médicamenteuses à base d'une ontologie OWL et de règles métier. In M.-C. JAULENT & L. SOUALMIA, Eds., *Actes de l'Atelier IC pour l'Interopérabilité Sémantique dans les applications en e-Santé*, Paris. https ://sites.google.com/site/icisante2012/programme.

CHOQUET R., MESSIAEN C., PRIOUZEAU A. & DE CARRARA A. (2012). Un jeu de données minimum pour faciliter l'interopérabilité des bases de données pour les maladies rares. In M.-C. JAULENT & L. SOUALMIA, Eds., *Actes de l'Atelier IC pour l'Interopérabilité Sémantique dans les applications en e-Santé*, Paris. https ://sites.google.com/site/icisante2012/programme.

COMBES P., COMBES S. & MONZIOLS M. (2016). Tentatives de suicide, prédire la récidive avec des techniques d'apprentissage statistique. In F. MOUGIN, A. ABDAOUI & P. ZWEIGENBAUM, Eds., *Actes du 2e Atelier IA & Santé*, Montpellier. https ://ic2016.sciencesconf.org/conference/ic2016/pages/IA_Sante.pdf.

COSSIN S., LOUSTAU R., JOUHET V., LÉTINIER L., MOUGIN F., EVRARD G., GIL-JARDINÉ C., DIALLO G. & THIESSARD F. (2018). ROMEDI, une terminologie médicale française pour la détection des médicaments en texte libre. In J. CHARLET, M.-D. DEVIGNES & B. SÉROUSSI, Eds., *Actes du 3e Atelier IA & Santé*, p. 17–22, Nancy. http ://pfia2018.loria.fr/communicationsiasante_3juillet2018/.

DALLOUX C., GRABAR N. & CLAVEAU V. (2017). Détection de la négation : corpus français et apprentissage supervisé. In L. TAMINE-LECHANI, Ed., *Actes de SIIM 2017, 4e Symposium sur l'Ingénierie de l'Information Médicale*, p. 17–24, Toulouse. https ://www.irit.fr/SIIM/2017/SIIM2017$_paper_17.pdf$.

DAYRE P. & BATATIA H. (2011). Annotation collaborative de médias pour l'émergence et l'apprentissage de concepts dans le milieu médical. In L. TAMINE, S. DARMONI & L. SOUALMIA, Eds., *Actes de SIIM 2011, 1er Symposium sur l'Ingénierie de l'Information Médicale*, p. 133–140, Toulouse. http ://www.irit.fr/SIIM/SIIM2011-12.pdf.

DE NIZZA D., ORTIZ J., MEURISSE H. & SCHOBBENS P.-Y. (2013). Formalisation et Construction d'une Ontologie dans le Domaine des Infections Orthopédiques. In R. TRONCY, Ed., *24èmes Journées Francophones d'Ingénierie des Connaissances - IC 2013*, Lille, France.

DEERWESTER S., DUMAIS S. T., FURNAS G. W., LANDAUER T. K. & HARSHMAN R. (1990). Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, **41**(6), 391–407.

DENDANI N. & ALLOUANI R. (2018). A decision support system for the diagnosis of the Diabetes disease. In J. CHARLET, M.-D. DEVIGNES & B. SÉROUSSI, Eds., *Actes du 3e Atelier IA & Santé*, p. 81–85, Nancy. http ://pfia2018.loria.fr/communicationsiasante_3juillet2018/.

DHOMBRES F. & CHARLET J. (2018). As ontologies reach maturity, Artificial Intelligence starts being fully efficient : Findings from the section on knowledge representation and management for the yearbook 2018. **27**(1), 140–145. https ://www.ncbi.nlm.nih.gov/pubmed/30157517.

DHOMBRES F., JOUANNIC J.-M., JAULENT M.-C. & CHARLET J. (2010). Choix méthodologiques pour la construction d'une ontologie de domaine en médecine prénatale. In S. DESPRÈS, Ed., *21èmes Journées Francophones d'Ingénierie des Connaissances - IC 2010*, p. 183–194, Nîmes, France : Ecole des Mines d'Alès.

DHOMBRES F., VANDENBUSSCHE P.-Y., RATH A., HANAEUR M., OLRY A., URBERO B., CHOQUET R. & CHARLET J. (2011). Projet OrphaOnto - Première étape de l'ontologisation des bases de connaissances d'Orphanet. In A. MILLE, Ed., *22èmes Journées Francophones d'Ingénierie des Connaissances - IC 2011*, p. 573–588, Chambéry, France.

DINH D., REIS J. C. D., SILVEIRA M. D. & PRUSKI C. (2013). Identification des informations conceptuelles définissant un alignement entre ontologies médicales. In L. TAMINE-LECHANI & L. SOUALMIA, Eds., *Actes de SIIM 2013, 2e Symposium sur l'Ingénierie de l'Information Médicale*, Lille. https ://www.irit.fr/SIIM/2013/05_siim13.pdf.

DINH D. & TAMINE L. (2011). Vers un modèle de recherche d'information multi-terminologique des documents biomédicaux. In L. TAMINE, S. DARMONI & L. SOUALMIA, Eds., *Actes de SIIM 2011, 1er Symposium sur l'Ingénierie de l'Information Médicale*, p. 43–56, Toulouse. http ://www.irit.fr/SIIM/SIIM2011-4.pdf.

DRAMÉ K., DIALLO G. & MOUGIN F. (2012). Construction d'une ontologie bilingue de la maladie d'Alzheimer à partir de textes médicaux. In M.-C. JAULENT & L. SOUALMIA, Eds., *Actes de l'Atelier IC pour l'Interopérabilité Sémantique dans les applications en e-Santé*, Paris. https ://sites.google.com/site/icisante2012/programme.

EHOLIÉ S., NZALI M. D. T., BRINGAY S. & JONQUET C. (2016). MuEVo, un vocabulaire multi-expertise (patient/médecin) dédié au cancer du sein. In F. MOUGIN, A. ABDAOUI & P. ZWEIGENBAUM, Eds., *Actes du 2e Atelier IA & Santé*, Montpellier. https ://ic2016.sciencesconf.org/conference/ic2016/pages/IA_Sante.pdf.

ESTEVA A., KUPREL B., NOVOA R. A., KO J., SWETTER S. M., BLAU H. M. & THRUN S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, **542**, 115.

FALIP J., BLANCHARD F. & HERBIN M. (2018). Exploration et système de recommandation pour l'aide au raisonnement médical. In J. CHARLET, M.-D. DEVIGNES & B. SEROUSSI, Eds., *Actes du 3e Atelier IA & Santé*, p. 87–91, Nancy. http ://pfia2018.loria.fr/communicationsiasante_3juillet2018/.

FIESCHI M., DING Y., TANG J., DONG X., HE B. & QIU J. (1990). The sphinx system environment. *Artificial Intelligence in Medicine*.

FRANDJI B. & JAULENT M.-C. (2012). Un SADC intégré et interopérable dans le Système d'Information Clinique. In M.-C. JAULENT & L. SOUALMIA, Eds., *Actes de l'Atelier IC pour l'Interopérabilité Sémantique dans les applications en e-Santé*, Paris. https ://sites.google.com/site/icisante2012/programme.

GAIGNARD A., MONTAGNAT J., ZUCKER C. F. & CORBY O. (2012). Fédération multi-sources en neu-

rosciences : intégration de données relationnelles et sémantiques. In M.-C. JAULENT & L. SOUALMIA, Eds., *Actes de l'Atelier IC pour l'Interopérabilité Sémantique dans les applications en e-Santé*, Paris. https ://sites.google.com/site/icisante2012/programme.

GARCIA-FERNANDEZ A., LIGOZAT A.-L. & BERNHARD D. (2011). Présent, hypothétique, conditionnel ? Annotation du statut des problèmes médicaux dans des comptes-rendus cliniques en français. In L. TAMINE, S. DARMONI & L. SOUALMIA, Eds., *Actes de SIIM 2011, 1erSymposium sur l'Ingénierie de l'Information Médicale*, p. 3–14, Toulouse. http ://www.irit.fr/SIIM/SIIM2011-1.pdf.

GHEZAIEL L. B., LATIRI C., AHMED M. B. & GOUIDER-KHOUJA N. (2011). Un réseau proxémique pour la recherche d'information : Application à la maladie des dystonies. In L. TAMINE, S. DARMONI & L. SOUALMIA, Eds., *Actes de SIIM 2011, 1er Symposium sur l'Ingénierie de l'Information Médicale*, p. 25–42, Toulouse. http ://www.irit.fr/SIIM/SIIM2011-3.pdf.

GRANDJEAN G. A., HOSSU G., BERTHOLDT C., NOBLE P., MOREL O. & GRANGÉ G. (2018). Artificial intelligence assistance for fetal head biometry : Assessment of automated measurement software. In J. CHARLET, M.-D. DEVIGNES & B. SÉROUSSI, Eds., *Actes du 3e Atelier IA & Santé*, p. 709–716, Nancy. http ://pfia2018.loria.fr/communicationsiasante_3juillet2018/.

GROSJEAN J., SOUALMIA L. F., MERABTI T., DAHAMNA B., KERGOURLAY I., THIRION B. & DARMONI S. J. (2011). The French health multi-terminology portal. In L. TAMINE, S. DARMONI & L. SOUALMIA, Eds., *Actes de SIIM 2011, 1er Symposium sur l'Ingénierie de l'Information Médicale*, p. 57–70, Toulouse. http ://www.irit.fr/SIIM/SIIM2011-5.pdf.

GUEFACK V. D., BERTAUD-GOUNOT V., BURGUN A., DUVAUFERRIER R. & LASBLEIZ J. (2012). Ontologie sémiologique biomédicale et sémiologie quantitative. In M.-C. JAULENT & L. SOUALMIA, Eds., *Actes de l'Atelier IC pour l'Interopérabilité Sémantique dans les applications en e-Santé*, Paris. https ://sites.google.com/site/icisante2012/programme.

HENRY V. J., SOUALMIA L. F., GROSJEAN J., DESFEUX A., DARMONI S. J. & GONZALEZ B. J. (2016). Omiconto : une ressource termino-ontologique pour la qualification et l'indexation des outils d'analyse en sciences omiques. In F. MOUGIN, A. ABDAOUI & P. ZWEIGENBAUM, Eds., *Actes du 2e Atelier IA & Santé*, Montpellier. https ://ic2016.sciencesconf.org/conference/ic2016/pages/IA_Sante.pdf.

HOFMANN T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, **42**(1-2), 177–196.

JANNOT A.-S., ZAPLETAL , BARBIERI A., GEROLDINGER A., BOULET S., ZOHAR S. & ZAANAN A. (2018). Intégration et synthèse visuelle de données multi- sources et hétérogènes de chimiothérapie. In J. CHARLET, M.-D. DEVIGNES & B. SEROUSSI, Eds., *Actes du 3e Atelier IA & Santé*, p. 23–29, Nancy. http ://pfia2018.loria.fr/communicationsiasante_3juillet2018/.

JONQUET C., COULET A., SHAH N. & MUSEN M. (2010). Indexation et intégration de ressources textuelles à l'aide d'ontologies : application au domaine biomédical. In S. DESPRÈS, Ed., *21èmes Journées Francophones d'Ingénierie des Connaissances - IC 2010*, p. 183–194, Nîmes, France : Ecole des Mines d'Alès.

KAMATH S., GRAU B. & MA Y. (2017). A Study of Word Embeddings for Biomedical Question Answering. In L. TAMINE-LECHANI, Ed., *Actes de SIIM 2017, 4e Symposium sur l'Ingénierie de l'Information Médicale*, p. 35–38, Toulouse. https ://www.irit.fr/SIIM/2017/SIIM2017$_paper_1$2.*pdf*.

KESSLER R., BÉCHET N., LEDEGEN G. & PUGNIERE-SAAVEDRA F. (2018). Exploration par apprentissage de discussions de personnes en détresse psychologique. In S. RANWEZ, Ed., *29es Journées Francophones d'Ingénierie des Connaissances, IC 2018*, p. 95–102, Nancy, France.

LAFOREST F., FLORY A. & GARIN-MICHAUD A. (2011). Accès transfrontalier aux informations médicales : un système de traduction pour le projet européen ALIAS. In L. TAMINE, S. DARMONI & L. SOUALMIA, Eds., *Actes de SIIM 2011, 1er Symposium sur l'Ingénierie de l'Information Médicale*, p. 5–23, Toulouse. http ://www.irit.fr/SIIM/SIIM2011-2.pdf.

LAMY J.-B., SOUALMIA L. F., VENOT A. & DUCLOS C. (2014). Validation de la sémantique d'un langage iconique médical à l'aide d'une ontologie : méthodes et applications. In C. FARON-ZUCKER, Ed., *25èmes Journées Francophones d'Ingénierie des Connaissances - IC 2014*, p. 51–62, Clermont-Ferrand, France. Session 1 : Construction, peuplement et exploitation d'ontologies.

LAMY J.-B., UGON A., DUCLOS C., VENOT A., FAVRE M. & BERTHELOT H. (2017). Une plate-forme visuelle pour une information comparative sur les nouveaux médicaments. In C. ROUSSEY, Ed., *28èmes Journées Francophones d'Ingénierie des Connaissances - IC 2017*, p. 38–49, Caen, France.

LAMY J.-B., UGON A., FAVRE M., VENOT A. & BERTHELOT H. (2015). Comparaison et visualisation des contre-indications des medicaments . In N. BRICON-SOUF, Ed., *Actes de SIIM 2015, 3e Symposium sur l'Ingénierie de l'Information Médicale*, p. 26–33, Rennes. https ://www.irit.fr/SIIM/2015/ActesSIIM2015.pdf.

LANDAUER T. K. & DUTNAIS S. T. (1997). A solution to plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *PSYCHOLOGICAL REVIEW*, **104**(2), 211–240.

LELONG R., SOUALMIA L., SAKJI S., DAHAMNA B. & DARMONI S. (2017). Une technologie NoSQL au service de moteur de recherche en santé. In L. TAMINE-LECHANI, Ed., *Actes de SIIM 2017, 4e Symposium sur l'Ingénierie de l'Information Médicale*, p. 41–48, Toulouse. https ://www.irit.fr/SIIM/2017/SIIM2017$_paper_2$.*pdf*.

LINDEMANN W. B. (2018). On the quality of automatic alerts during electronic prescription and the possibilities of improvement. In J. CHARLET, M.-D. DEVIGNES & B. SÉROUSSI, Eds., *Actes du 3e Atelier IA & Santé*, p. 93–99, Nancy. http ://pfia2018.loria.fr/communicationsiasante_3juillet2018/.

MAAROUFI M., CHOQUET R. & LANDAIS P. (2014). Formalisation des correspondances pour l'optimisation des alignements automatisés de schémas de données : Application au domaine des maladies rares.

In S. BRINGAY, N. SOUF & L. TAMINE-LECHANI, Eds., *Actes de l'Atelier IC Santé*, Clermont-Ferrand. https ://www.lirmm.fr/ic-sante/pmwiki/docs/Maaroufi.pdf.

MARRAST P., SOUF N. & ZARATÉ P. (2013). Heuristiques et annotations pour (re)penser le Dossier Clinique Informatisé. In L. TAMINE-LECHANI & L. SOUALMIA, Eds., *Actes de SIIM 2013, 2e Symposium sur l'Ingénierie de l'Information Médicale*, Lille. https ://www.irit.fr/SIIM/2013/01_siim13.pdf.

MARTIN L., BATTISTELLI D. & CHARNOIS T. (2014). Mise en place d'une méthode de reconnaissance des signes et des symptômes dans le contexte des maladies rares. In S. BRINGAY, N. SOUF & L. TAMINE-LECHANI, Eds., *Actes de l'Atelier IC Santé*, Clermont-Ferrand. https ://www.lirmm.fr/ic-sante/pmwiki/docs/Martin.pdf.

MARY M., SOUALMIA L. F. & GANSEL X. (2016). Interopérabilité sémantique dans le domaine du diagnostic in vitro. In F. MOUGIN, A. ABDAOUI & P. ZWEIGENBAUM, Eds., *Actes du 2e Atelier IA & Santé*, Montpellier. https ://ic2016.sciencesconf.org/conference/ic2016/pages/IA_Sante.pdf.

MEILENDER T., JAY N., LIEBER J. & PALOMARES F. (2011). Édition sémantique d'arbres de décision pour l'oncologie avec KCATOS. In L. TAMINE, S. DARMONI & L. SOUALMIA, Eds., *Actes de SIIM 2011, 1er Symposium sur l'Ingénierie de l'Information Médicale*, p. 121–132, Toulouse. http ://www.irit.fr/SIIM/SIIM2011-11.pdf.

MERABTI T., SOUALMIA L. F., GROSJEAN J., JOUBERT M. & DARMONI S. J. (2012). Méthodes d'alignement de terminologies médicales et leur intégration dans un portail. In M.-C. JAULENT & L. SOUALMIA, Eds., *Actes de l'Atelier IC pour l'Interopérabilité Sémantique dans les applications en e-Santé*, Paris. https ://sites.google.com/site/icisante2012/programme.

MERCADIER Y., AZÉ J., BRINGAY S., CLAVIER V., CUENCA E., PAGANELLI C., PONCELET P. & SALLABERRY A. (2018). #AIDS Analyse Information Dangers Sexualité : caractériser les discours à propos du VIH dans les forums de santé. In S. RANWEZ, Ed., *29èmes Journées Francophones d'Ingénierie des Connaissances - IC 2018*, p. 71–86, Nancy, France : AFIA.

MINARD A.-L., LIGOZAT A.-L. & GRAU B. (2011). Extraction de relations dans des comptes rendus hospitaliers. In A. MILLE, Ed., *22èmes Journées Francophones d'Ingénierie des Connaissances - IC 2011*, p. 491–506, Chambéry, France.

MOULAHI B., AZÉ J. & BRINGAY S. (2017). Suivi et détection des idéations suicidaires dans les médias sociaux. In C. ROUSSEY, Ed., *28èmes Journées Francophones d'Ingénierie des Connaissances - IC 2017*, p. 26–37, Caen, France.

NEURAZ A., GARCELON N., BURGUN A. & RANCE B. (2018). multiWAS : interactive and multimodal phenome-wide scan for data-driven cohort description. In J. CHARLET, M.-D. DEVIGNES & B. SEROUSSI, Eds., *Actes du 3e Atelier IA & Santé*, p. 31–36, Nancy. http ://pfia2018.loria.fr/communicationsiasante_3juillet2018/.

NGUYEN G.-H., TAMINE L., SOULIER L. & SOUF N. (2017). Apprentissage de représentation des documents médicaux guidé par les concepts pour la recherche d'information. In L. TAMINE-LECHANI, Ed., *Actes de SIIM 2017, 4e Symposium sur l'Ingénierie de l'Information Médicale*, p. 27–34, Toulouse. https ://www.irit.fr/SIIM/2017/SIIM2017$_paper_8.pdf$.

NIKIEMA J. N., JOUHET V. & MOUGIN F. (2016). Evaluation de la SNOMED CT comme support à l'alignement de terminologies diagnostiques en cancérologie. In F. MOUGIN, A. ABDAOUI & P. ZWEIGENBAUM, Eds., *Actes du 2e Atelier IA & Santé*, Montpellier. https ://ic2016.sciencesconf.org/conference/ic2016/pages/IA_Sante.pdf.

NIKIEMA J. N., MOUGIN F. & JOUHET V. (2017). Processus de prétraitement des libellés d'une terminologie d'interface. In L. TAMINE-LECHANI, Ed., *Actes de SIIM 2017, 4e Symposium sur l'Ingénierie de l'Information Médicale*, p.9̃, Toulouse.

NOTARO M., SCHUBACH M., ROBINSON P. N. & VALENTINI G. (2017). Prediction of Human Phenotype Ontology terms by means of hierarchical ensemble methods. *BMC Bioinformatics*, **18**(1).

OPITZ T., AZÉ J., BRINGAY S., JOUTARD C. & MOLLEVI C. (2014). Paroles de patients dans les forums de santé : une perspective originale sur la qualité de la vie. In S. BRINGAY, N. SOUF & L. TAMINE-LECHANI, Eds., *Actes de l'Atelier IC Santé*, Clermont-Ferrand. https ://www.lirmm.fr/ic-sante/pmwiki/docs/Opitz.pdf.

PARÈS Y., AIMÉ X., CHARLET J. & JAULENT M.-C. (2014). Vers une harmonisation automatique de la représentation de comptes rendus medicaux pour évaluer leurs similarités. In S. BRINGAY, N. SOUF & L. TAMINE-LECHANI, Eds., *Actes de l'Atelier IC Santé*, Clermont-Ferrand. https ://www.lirmm.fr/ic-sante/pmwiki/docs/Pares.pdf.

PASCO J., CAMPILLO-GIMENEZ B., GRAMMATICO-GUILLON L. & CUGGIA M. (2018). Pré-screening en cancérologie : automatisation à partir des entrepôts de données cliniques. In J. CHARLET, M.-D. DEVIGNES & B. SEROUSSI, Eds., *Actes du 3e Atelier IA & Santé*, p. 37–42, Nancy. http ://pfia2018.loria.fr/communicationsiasante_3juillet2018/.

PERSONENI G., BRESSO E., DEVIGNES M.-D., SMAÏL-TABBONE M. & COULET A. (2018). Découverte d'associations entre Evénements Indésirables Médicamenteux par les structures de patrons et les ontologies. In J. CHARLET, M.-D. DEVIGNES & B. SEROUSSI, Eds., *Actes du 3e Atelier IA & Santé*, p. 43–48, Nancy. http ://pfia2018.loria.fr/communicationsiasante_3juillet2018/.

PERSONENI G., DEVIGNES M.-D., DUMONTIER M., SMAÏL-TABBONE, MALIKA M. & COULET, ADRIEN A. (2016). Extraction d'associations d'EIM à partir de dossiers patients : expérimentation avec les structures de patrons et les ontologies. In F. MOUGIN, A. ABDAOUI & P. ZWEIGENBAUM, Eds., *Actes du 2e Atelier IA & Santé*, Montpellier. https ://ic2016.sciencesconf.org/conference/ic2016/pages/IA_Sante.pdf.

PERTIN C., DECCACHE C., GAGNAYRE R. & HAMON T. (2017). Identication du profil des utilisateurs dans les

forums de discussion de santé. In L. TAMINE-LECHANI, Ed., *Actes de SIIM 2017, 4e Symposium sur l'Ingénierie de l'Information Médicale*, p. 85–92, Toulouse. https ://www.irit.fr/SIIM/2017/SIIM2017$_p aper_1 8.pdf$.

PETEGROSSO R., PARK S., HWANG T. H. & KUANG R. (2016). Transfer learning across ontologies for phenome–genome association prediction. *Bioinformatics*, p. btw649.

PINAIRE J., ALOUANE S. B., AZÉ J., BRINGAY S., LANDAIS P. & SALLABERRY A. (2015a). Visualisation interactive de trajectoires de patients. Poster.

PINAIRE J., AZÉ J., BRINGAY S. & LANDAIS P. (2016). Extraire semi-automatiquement des connaissances dans la littérature biomédicale . In N. PERNELLE, Ed., *27èmes Journées Francophones d'Ingénierie des Connaissances - IC 2016*, Montpellier, France.

PINAIRE J., AZÉ J., BRINGAY S. & LANDAIS P. (2017a). Infarctus du myocarde : quelles sont les trajectoires de soins pronostiques du décès à l'hôpital ? In C. ROUSSEY, Ed., *28èmes Journées Francophones d'Ingénierie des Connaissances - IC 2017*, p. 14–25, Caen, France.

PINAIRE J., AZÉ J., BRINGAY S., PONCELET P., GENOLINI C. & LANDAIS P. (2017b). Quels évènements après un infarctus du myocarde ? In L. TAMINE-LECHANI, Ed., *Actes de SIIM 2017, 4e Symposium sur l'Ingénierie de l'Information Médicale*, p. 69–76, Toulouse. https ://www.irit.fr/SIIM/2017/SIIM2017_paper_4.pdf.

PINAIRE J., RABATEL J., AZÉ J. & BRINGAY S. (2015b). Recherche et visualisation de trajectoires dans les parcours de soins des patients ayant eu un infarctus du myocarde. In N. BRICON-SOUF, Ed., *Actes de SIIM 2015, 3e Symposium sur l'Ingénierie de l'Information Médicale*, p. 5–12, Rennes. https ://www.irit.fr/SIIM/2015/ActesSIIM2015.pdf.

PSIUK T. & MANET P. (2015). Intégrer les chemins cliniques dans les outils de soin : rêve ou réalité ? In N. BRICON-SOUF, Ed., *Actes de SIIM 2015, 3e Symposium sur l'Ingénierie de l'Information Médicale*, p. 13–18, Rennes. https ://www.irit.fr/SIIM/2015/ActesSIIM2015.pdf.

RABOUDI A., ALLANIC M., HERVÉ P.-Y., BOUTINAUD P., DURUPT A., BALVAY D. & EYNARD B. (2017). Traçabilité de l'intégration de données biomédicales hétérogènes dans le système SWOMed de gestion du cycle de vie des études biomédicales. In L. TAMINE-LECHANI, Ed., *Actes de SIIM 2017, 4e Symposium sur l'Ingénierie de l'Information Médicale*, p. 57–65, Toulouse. https ://www.irit.fr/SIIM/2017/SIIM2017_paper_13.pdf.

RANWEZ S., HARISPE S., CHARLET J., PICARD A.-C. L., JEAN P. & MÉNAGER M.-T. (2015). OntoToxNuc : exploitation d'une ontologie de la toxicologie nucléaire environnementale dans une plateforme collaborative. In N. BRICON-SOUF, Ed., *Actes de SIIM 2015, 3e Symposium sur l'Ingénierie de l'Information Médicale*, p. 19–25, Rennes. https ://www.irit.fr/SIIM/2015/ActesSIIM2015.pdf.

REIS J. C. D., PRUSKI C. & SILVEIRA M. D. (2012). Vers une approche automatique pour la maintenance des mappings entre ressources termino- ontologiques du domaine de la santé. In M.-C. JAULENT & L. SOUALMIA, Eds., *Actes de l'Atelier IC pour l'Interopérabilité Sémantique dans les applications en e-Santé*, Paris. https ://sites.google.com/site/icisante2012/programme.

RICHARD A., MAYAG B., MEINARD Y., TALBOT F. & TSOUKIAS A. (2018). How AI could help physicians during their medical consultations : An analysis of physicians' decision process to develop efcient decision support systems for medical consultations. In J. CHARLET, M.-D. DEVIGNES & B. SEROUSSI, Eds., *Actes du 3e Atelier IA & Santé*, p. 100–106, Nancy. http ://pfia2018.loria.fr/communicationsiasante_3juillet2018/.

RICHARD M., AIMÉ X., KREBS M.-O. & CHARLET J. (2013). L'Ingénierie des Connaissances à l'usage du PMSI en Psychiatrie. In L. TAMINE-LECHANI & L. SOUALMIA, Eds., *Actes de SIIM 2013, 2e Symposium sur l'Ingénierie de l'Information Médicale*, Lille. https ://www.irit.fr/SIIM/2013/04_siim13.pdf.

ROSENBLATT F. (1957). The perceptron—a perceiving and recognizing automaton. *Cornell Aeronautical Laboratory*, p. 85–460–1.

ROUSSEY C. & GHORFI T. A. (2018). Annotation sémantique pour une interrogation experte des Bulletins de Santé du Végétal. In S. RANWEZ, Ed., *29èmes Journées Francophones d'Ingénierie des Connaissances - IC 2018*, p. 37–52, Nancy, France : AFIA.

RYBARCZYK Y., MARTINS R. & FONSECA J. (2011). Plateforme de traitement multimodal des symptômes aphasiques. In L. TAMINE, S. DARMONI & L. SOUALMIA, Eds., *Actes de SIIM 2011, 1er Symposium sur l'Ingénierie de l'Information Médicale*, p. 141–150, Toulouse. http ://www.irit.fr/SIIM/SIIM2011-8.pdf.

SCHNELL M., COUFGNAL S., LIEBER J., SALEH S. & JAY N. (2018). Interprétation de bonnes pratiques de codication médicale par du raisonnement à partir de cas — Application à la saisie de données pour les registres du cancer. In J. CHARLET, M.-D. DEVIGNES & B. SEROUSSI, Eds., *Actes du 3e Atelier IA & Santé*, p. 49–54, Nancy. http ://pfia2018.loria.fr/communicationsiasante_3juillet2018/.

SEDIRI M., MATTA N., LORIETTE S. & HUGEROT A. (2012). Vers une représentation de situations de crise gérées par le SAMU. In S. SZULMAN, Ed., *23èmes Journées Francophones d'Ingénierie des Connaissances - IC 2012*, Paris, France.

SEGOND F., PONOMAREVA A., RABARIJAONA D., DINI L., KERGOURLAY I., DARMONI S., GICQUEL Q. & METZGER M. H. (2014). Bien représenter pour mieux raisonner : deux approches pour le dossier patient. In S. BRINGAY, N. SOUF & L. TAMINE-LECHANI, Eds., *Actes de l'Atelier IC Santé*, Clermont-Ferrand. https ://www.lirmm.fr/ic-sante/pmwiki/docs/Segond.pdf.

SERRES B., ZEMMOURA I., DESTRIEUX C. & VENTURINI G. (2011). Acquisition, visualisation 3d et interactions pour le suivi de dissection. In L. TAMINE, S. DARMONI & L. SOUALMIA, Eds., *Actes de SIIM 2011, 1er Symposium sur l'Ingénierie de l'Information Médicale*, p. 103–110, Toulouse. http ://www.irit.fr/SIIM/SIIM2011-9.pdf.

SEYMOUR K. & PAYOUX P. (2017). Radiomics Enabler®, an ETL (Extract-Transform-Load) for biomedical imaging in big-data projects. In L. TAMINE-LECHANI, Ed., *Actes de SIIM 2017, 4e Symposium sur l'Ingénie-*

rie de l'Information Médicale, p. 49–55, Toulouse. https ://www.irit.fr/SIIM/2017/SIIM2017_paper_5.pdf.

SOW A., GUISSÉ A. & NIANG O. (2018). Intégration d'ontologies médicales : amélioration par association des maladies humaines à leurs plus pertinents signes caratéristiques. In S. RANWEZ, Ed., *29èmes Journées Francophones d'Ingénierie des Connaissances - IC 2018*, Nancy, France : AFIA.

SUN R. (1999). Artificial intelligence : Connectionist and symbolic approaches.

SÉROUSSI B., GALOPIN A. & GAOUAR M. (2018). Utilisation des cercles thérapeutiques pour l'affichage des recommandations de bonne pratique dans la prise en charge des patients polypathologiques. In J. CHARLET, M.-D. DEVIGNES & B. SEROUSSI, Eds., *Actes du 3e Atelier IA & Santé*, p. 55–61, Nancy. http ://pfia2018.loria.fr/communicationsiasante_3juillet2018/.

SÉROUSSI B., MESSAI N., LAOUÉNAN C., MENTRÉ F. & BOUAUD J. (2013). Profils patients associés à la non conformité des décisions aux recommandations de prise en charge thérapeutique des cancers du sein : utilisation de l'analyse de concepts formels. In R. TRONCY, Ed., *24èmes Journées Francophones d'Ingénierie des Connaissances - IC 2013*, Lille, France.

SÉROUSSI B., SAUQUET D., FALCOFF H., JULIEN J. & BOUAUD J. (2011). Formalisation de l'attitude des médecins vis à vis des propositions d'un système d'aide à la décision : évaluation de l"' e-iatrogénie " sur un cas d'hypertension avec ASTI mode guidé. In A. MILLE, Ed., *22èmes Journées Francophones d'Ingénierie des Connaissances - IC 2011*, p. 673–688, Chambéry, France.

TAPI NZALI M. D., BRINGAY S., LAVERGNE C., OPITZ T., AZÉ J. & MOLLEVI C. (2015). Construction d'un vocabulaire patient/médecin dédié au cancer du sein à partir des médias sociaux. In M.-H. ABEL, Ed., *26èmes Journées Francophones d'Ingénierie des Connaissances - IC 2015*, Rennes, France.

TERNOIS I., ESCUDIÉ J. B. & DUCLOS C. (2018). Développement et évaluation d'une méthode de codage automatique des endoscopies digestives. In J. CHARLET, M.-D. DEVIGNES & B. SEROUSSI, Eds., *Actes du 3e Atelier IA & Santé*, p. 63–68, Nancy. http ://pfia2018.loria.fr/communicationsiasante_3juillet2018/.

THEBAULT M., AMARDEILH F., DELAMARRE D., GUILLEMIN-LANNE S., JAMET A. & LILLO-LE LOUET A. (2010). VigiTermes : une plateforme de recherche et d'analyse des publications scientifiques au service de la pharmacovigilance. In S. DESPRÈS, Ed., *21èmes Journées Francophones d'Ingénierie des Connaissances - IC 2010*, p. 183–194, Nîmes, France : Ecole des Mines d'Alès.

TIRILLY P. (2011). Contributions à l'usage de la modalité dans les systèmes de recherche d'images médicales. In L. TAMINE, S. DARMONI & L. SOUALMIA, Eds., *Actes de SIIM 2011, 1erSymposium sur l'Ingénierie de l'Information Médicale*, p. 87–102, Toulouse. http ://www.irit.fr/SIIM/SIIM2011-8.pdf.

TRAORE L., CHNITI A., HUSSAIN S., GRIFFON N., DARMONI S., CHARLET J., SADOU E., OUAGNE D., LEPAGE E. & DANIEL C. (2014). Plateforme d'interopérabilité sémantique gérant les terminologies d'interface au sein d'un espace de partage. In C. FARON-ZUCKER, Ed., *25èmes Journées Francophones d'Ingénierie des Connaissances - IC 2014*, p. 75–86, Clermont-Ferrand, France. Session 1 : Construction, peuplement et exploitation d'ontologies.

UGON A. (2018). De la difculté d'évaluer les algorithmes de scorage automatique des données médicales temporelles : exemple de la polysomnographie. In J. CHARLET, M.-D. DEVIGNES & B. SEROUSSI, Eds., *Actes du 3e Atelier IA & Santé*, p. 69–74, Nancy. http ://pfia2018.loria.fr/communicationsiasante_3juillet2018/.

UGON A., KOTTI A., SEDKI K., PHILIPPE C., SÉROUSSI B., BOUAUD J., GANASCIA J.-G., GARDA P. & PINNA A. (2016). Reconnaissance des stades de sommeil à l'aide d'un outil de support à la décision basé sur les connaissances et la pratique des experts. In F. MOUGIN, A. ABDAOUI & P. ZWEIGENBAUM, Eds., *Actes du 2e Atelier IA & Santé*, Montpellier. https ://ic2016.sciencesconf.org/conference/ic2016/pages/IA_Sante.pdf.

WANG H., DING Y., TANG J., DONG X., HE B. & QIU J. (2011). Finding complex biological relationships in recent pubmed articles using bio-lda. *PLoS One*, **6**(3), 69.

WANG S.-H., DING Y., ZHAO W., HUANG Y.-H., PERKINS R., ZOU W. & CHEN J. J. (2016). Text mining for identifying topics in the literatures about adolescent substance use and depression. *BMC Public Health*, **16**(1), 279.

ZNAIDI E., TAMINE L., CHOUQUET C. & LATIRI C. (2013). Analyse exploratoire des requêtes d'experts médicaux : cas des campagnes d'évaluation TREC et CLEF. In L. TAMINE-LECHANI & L. SOUALMIA, Eds., *Actes de SIIM 2013, 2e Symposium sur l'Ingénierie de l'Information Médicale*, Lille. https ://www.irit.fr/SIIM/2013/02_siim13.pdf.

Recent progresses in data and knowledge integration for decision support in agri-food chains

Sophie Aubin[1], Pierre Bisquert[2], Patrice Buche[2], Juliette Dibie[3], Liliana Ibanescu[3], Clément Jonquet[4], Catherine Roussey[5]
[1]IST INRA, Angers, France
sophie.aubin@inra.fr
[2]IATE INRA, INRIA GraphiK, Montpellier University, France
{pierre.bisquert, patrice.buche}@inra.fr
[3]UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005 Paris, France
{dibie,liliana.ibanescu}@agroparistech.fr
[4]LIRMM, Université de Montpellier & CNRS, France
jonquet@lirmm.fr
[5]Université Clermont Auvergne, Irstea, UR TSCF, Centre de Clermont-Ferrand,
9 avenue Blaise Pascal CS 20085, F-63178 Aubière, France
catherine.roussey@irstea.fr

Abstract : Research in Agri-food and related fields dealing with sustainability has undergone important changes in the past years and tends to be more integrative, collaborative, and interdisciplinary. In this article, we present an excerpt of recent and ongoing projects in the French community proposing data integration guided by an ontology for collective decision support in the Agri-food domain. The use of ontologies for primary production and secondary production, i.e. transformation of primary production into food products, are addressed. An example of collective decision support system for food packaging selection which relies on scientific data annotated thanks to ontologies and food chain stakeholders' preferences is described. French initiatives at the international level to share agri-food ontologies are presented.

Key-words : Ontology, data integration, flexible querying, computational social choice, argumentation

1 Introduction

Research in Agri-food and related fields dealing with sustainability has undergone important changes in the past years and tends to be more integrative, collaborative, and interdisciplinary (Perrot et al., 2016). The Agri-food domain is considered as an interconnected system with various entities and complex relationships among them (Wolfert et al., 2010). More and more numerous data coming from heterogeneous sources cover the whole food chain and can be combined to address new questions. For instance, to test a hypothesis about the effects of different viticulture treatments on wine quality, researchers need to access and analyze various data sources at different scales, from the field to the bottle. Data integration is not so easy and researchers have to face several issues. Data are commonly stored in scattered places and their formats, naming, storage and query or retrieval mechanisms are very diverse. The heterogeneity of scientific data may come from many factors, such as (i) they are collected separately based on independent research projects with very specific targets and aims; (ii) the data structure frequently depends upon the collection method (e.g., to make data easier to record) or is function of future analysis, instead of using standard data representation (e.g., relational database schema); (iii) the terms and concepts used to annotate data are not standardized, neither within nor across scientific disciplines and research groups (Bowers, 2012). The difficulties in organizing data and knowledge in a unified way do not only limit research productivity but also reduce data traceability (Gardner, 2005). Research experiments are commonly divided into sub-domains, such as agricultural production, post-harvest, and food transformation. Even though the explicit links between those sub-domains may be easy to explicit and understand, each of them has different objectives, scopes and circumstances. For instance, agricultural experiments are usually conducted in the plots where environmental factors are difficult to control, while food processing experiments are generally carried out in

laboratories with controlled environmental conditions. From a practical point of view, both experiments require different methods for collecting and organizing observational data that yield differences in data format, structure and storage. The heterogeneity also occurs due to the vast scope of Agri-food sub-domains. For instance, each discipline and sub-domain involved in Agri-food domains uses its own way to express knowledge, terms, concepts and semantic relationships, which makes the scientific data sharing harder.

Studies in the last two decades have shown that ontologies represent a flexible way to link the information contained in heterogeneous data sources within or across domains (Gardner, 2005; Seedah et al., 2015; Doan et al., 2012). An ontology defines a set of primitives to model a domain of interest: classes, attributes (or properties) and relations between members of the classes (Guarino et al., 2009). Ontology is used to create and/or reuse standardized vocabularies and to index data sources with those vocabularies in order to allow data source interoperability. It opens the possibility to draw more comprehensive conclusions and to view data from different perspectives. Ontologies also allow certain types of automated reasoning to be performed. These features will help to develop advanced Information Systems able to manage heterogeneous data sources and to design platforms for more collaborative scientific data analysis and decision support to food chains.

In this article, we present an excerpt of recent and ongoing projects in the French community proposing data integration guided by an ontology for collective decision support in the Agri-food domain. Section 2 addresses the use of ontologies for primary production. Section 3 presents the use of ontologies for secondary production, i.e. transformation of primary production into food products. An example of decision support system for food packaging selection which relies on scientific data annotated thanks to ontologies will be described in the Section 4. Section 5 presents French initiatives at the international level to share agri-food ontologies. Finally, we conclude in Section 6.

2 Data integration guided by ontology for primary production

Precision agriculture is nowadays getting more and more attention in Europe due to the development of remote sensors and wireless sensors. Precision agriculture can be summarized as putting the right amount of matter (e.g., water, nutriment or pesticide) at the right time and at the right place. The goal is to reduce the quantity of matter that are lost in the soil or the air. Precision agriculture is based on accurate observations. These observations come from various types of sensors such as farmers (i.e., human sensors), remote sensors or sensors in the field. These observations could be quite complex. Some precise protocols should be followed to perform an accurate observation.

In this section we focus on crop observations, that are observations made on cultivated plots. These observations can be for instance the crop growth stage, the presence of pests in the cultivated plot, observations of pedo climatic conditions (e.g., soil moisture, quantity of rain, outdoor temperature). These observations help identifying potential risks like water stress. For a long time, agronomists have developed models (e.g., simulation, decision tree) to evaluate these risks and to help the design of Decision Support Systems (DSS) consuming crop observations. DSS may activate different types of actions: send alerts to farmers' smartphone in order to advise them to consolidate the risk evaluation by observing the presence of the problems on the plot. DSS may also automatically activate an equipment to solve the problem (e.g., irrigate a plot if the water stress risk is high).

Due to the fact that all these observations come from various kinds of sensors and have many different types, the storage of these observations suffers from heterogeneity issues. One solution is to use ontologies to store all these observations. The focus of Section 2 is to present the advantages of using a well-known ontology design pattern (Property and FeatureOfInterest)

to store in a similar way all types of observations. In Section 2.3 we present two uses cases developed at Irstea, a French Research Institute for Agriculture.

2.1 Ontologies related to Observations

From our point of view, there are two types of observations in primary production: (1) observations made by sensor located in the cultivated plot, such as soil moisture probes or agricultural weather stations; (2) observations made by farmers or any crop experts following an observation protocol. Ontologies for storing sensor measurements were very well studied in the literature. We can cite for example Semantic Sensor Network (SSN) (Compton et al., 2012) and Smart Appliances REFerence For Environment (SAREF4ENVI) (ETSI, 2017). Concerning human observations, the Extensible Observation Ontology (OBOE) (Madin et al., 2007) is the only one dedicated to environmental scientific observations. Note that the ontologies dedicated to the description of experiments are not in the scope of this section.

The SSN ontology is used as the core of the crop observation model and more precisely a new module, call SOSA.

SSN is an ontology developed by the Semantic Sensor Network Incubator Group (SSN-XG). The final report of the original SSN was published on the site of World Wide Web Consortium (W3C) on 28th June 2011. SSN ontology describes sensors and their measurements (Compton et al., 2012). SSN ontology defines a ***Sensor*** as "*Device, agent (including humans), or software (i.e. simulation) involved in, or implementing, a Procedure. Sensors respond to a Stimulus, e.g., a change in the environment, or Input data composed of the Results of prior Observations, and generate a Result…*" .

The joint W3C and OGC Spatial Data on the Web Working Group has developed a new version of the SSN including a module called SOSA (Sensor, Observation, Sampler and Actuator) (Haller et al., 2017). This module replaces the SSO (Stimulus Sensor Observation) Pattern. To describe sensing acts, SOSA provides ***sosa:Sensor*** (e.g. a thermometer) that make ***sosa:Observation*** (e.g. a measurement) about ***sosa:ObservableProperty*** (e.g. temperature). The s***osa:ObservableProperty*** (e.g. temperature) is a property of a ***sosa:FeatureOfInterest*** (e.g. the outdoor air, the air of a specific room, a human body, an oven). Moreover, to identify the specimen where the sensing act was performed, SOSA presents ***sosa:Sampler*** that makes ***sosa:Sampling*** of some ***sosa:FeatureOfInterest*** to produce a ***sosa:Sample***. The ***sosa:Sample*** that corresponds to the ***sosa:FeatureOfInterest*** air is the volume of air around the weather station. This design pattern is useful to describe precisely the sensor measurement. For example our weather station interface presents its temperature measurement with the label "temperature". Everybody will understand that the weather station measures the outdoor temperature of the air. But other thermometer may be used in agricultural equipments: soil, building or cattle. So the precision of the SSN/SOSA design pattern enables to clarify the sensor measurements and it is the core of the model detailed in the next section.

2.2 Crop Observation Model based on SSN/SOSA

To build our model of crop observation based on SSN/SOSA we needed others ontologies presented in Table 1. The acronyms given in the second column of Table 1 are used in Figure 1.

TABLE 1 - *the list of ontologies used to model crop observation*

Ontology name	**acronym**	**authors**	**reference**
Semantic Sensor Network	sosa	W3C	(Haller et al, 2017)
Time	time	W3C	(Hobbs & Pan, 2004)
GeoSparql	geo	OGC	(Battle & Kolas, 2012)
Prov ontology	prov	W3C	(Lebo et al., 2013)
Simple Knowledge Organisation System	skos	W3C	
Quantities, Units, Dimensions and data Types	qudt	QUDT.org member of W3C	(Hodgson et al., 2014)

As shown in Figure 1, the crop observation denoted `P25Maize20170101` is an instance of the class ***sosa:Observation***. This observation is related to a cultivated plot, denoted `P25` in Figure 1. A cultivated plot is modeled as an instance of ***sosa:Sample*** because a cultivated plot at a specific time is indeed a specific sample of a crop (e.g. maize). Therefore the crop is an instance of ***sosa:FeatureOfInterest*** which has as many samples as cultivated plots of the crop exist. The plot is also a geographic object. So the representation of a cultivated plot is an instance of both ***sosa:Sample*** and of ***geo:Feature*** classes.

The properties used to describe the observation are:

- ***sosa:hasFeatureOfInterest*** that links an instance of ***sosa:Observation*** to an instance of ***sosa:Sample***. The instance of a sample may be one specific cultivated plot or a set of plots depending on the use case.
- ***sosa:isSampleOf*** that links the instance representing the plot to an instance of ***skos:Concept*** representing the type of crop (e.g. maize). This skos instance may come from any thesaurus describing crop like Agrovoc (http://aims.fao.org/fr/agrovoc).
- ***sosa:observedProperty*** that links an instance of ***sosa:Observation*** to an instance of ***skos:Concept*** representing the crop growth stage, denoted `V2` in Figure 1. This skos instance may come from any thesaurus. For example we can reuse the CROP ontology (http://agroportal.lirmm.fr/ontologies/CO_715) that contains the BBCH crop growth stage, a scale for a uniform coding of growth stages.
- ***sosa:phenomenonTime*** that links an instance of ***sosa:Observation*** to an instance of ***time:Instant*** or ***time:Interval***. In the case of an observation of crop growth stage the time entity is an interval that represent the period when the cultivated plot reach the growth stage. Note that the interval is described by several time entity like the beginning instant as presented in Figure 1.
- ***sosa:resultTime*** that links an instance of ***sosa:Observation*** to an ***xsd:dateTime*** value that represents the day when the observation was done.
- ***sosa:hasSimpleResult*** is an attribute that contains a percentage. This percentage may express different ratio depending of the use case. It may be the percentage of plants that reach the given growth stage on the number of plants in the plot. Note that a plot reaches a given growth stage when 50% of the plants of the plot reach this growth stage. Other interpretation of the percentage may be the ratio between the number of plots that has reached the crop growth stage on the total number of plots in the sample.
- ***sosa:hasResult*** that links an instance of ***sosa:Observation*** to an instance of ***sosa: Result.*** This instance has two attributes: one indicates the unit of the measurement and the other the value.

- ***sosa:madeBySensor*** that links an instance of ***sosa:Observation*** to an instance of ***sosa:Sensor***. This instance may be a person or a device depending of the use case.

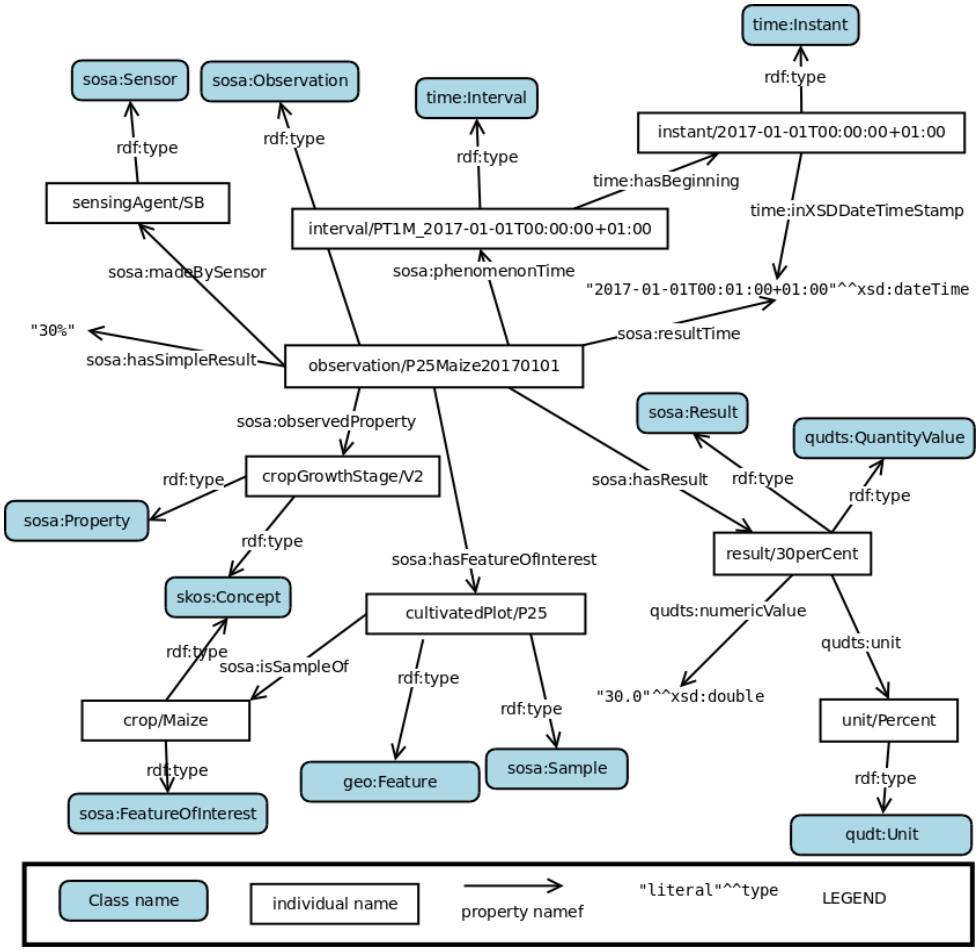


FIGURE 1 - *an example of crop observation based on SSN/SOSA*.

2.3 Two Use Case Descriptions in Primary Production

The crop model observation described in the previous section is implemented in two use cases: the design of an archive of crop observations and the design of a context aware system to automate irrigation.

The first use case is part of the French ANR D2KAB project. The goal of this project is to build a search engine dedicated to agricultural alert bulletins called “Bulletin de Santé du Végétal”. An alert bulletin is dedicated to a specific area (e.g. a French administrative unit) and a specific crop category (e.g. cereals). A bulletin contains some human observations: crop growth stage and pest presence. It also contains some evaluation of risk about pest attacks. Note that a presence of a pest on a cultivated plot becomes an attack only if the crop has reached a defined growth stage and if the presence of the pest is significant. Some protocols of pest observation are defined to evaluate the pest attack. Otherwise the pest has no impact on the crop production.

A bulletin does not contain an aggregation of all observations about one crop, but it contains few useful observations selected by specialists, the authors of the bulletin. One of the goals of the D2KAB project is to extract from the text content of the bulletin the crop growth

stage observations. Results are published in (Roussey & Abderrahmani Ghorfi 2018). The outcome of this project is that the annotated corpus becomes a spatio temporal archive of crop observations for France and may be queried for different purposes.

The second use case designs a context-aware system to automate maize irrigation. The context of a cultivated plot is acquired by a wireless sensor network located on this plot. This context contains some automatic evaluation of crop growth stage based on temperature measurements. It also contains some soil moisture measurements and pluviometer measurements. In order to take an irrigation decision, the context aware system stores different types of observations: raw sensor measurements, spatio-temporal aggregation of measurements, qualitative data inferred from the comparison of the aggregated value and a threshold. A decision process starts from some simple quantitative measurements (e.g., soil moisture at 30 cm depth is 42 cbar) and finishes by a qualitative data that summarize the crop situation (e.g., crop water stress is high). Results are published in (Poveda et al, 2018) and confirm that the crop water stress depend on its growth stage and the amount of water in the soil.

3. Data integration guided by ontology for secondary production

One of the most significant current discussion in food production is to formulate food products with high qualities such as nutritional requirements or sensory acceptable by consumers as well as low environmental impact. This supposes to build a decision support tool combining data and knowledge from different domains in food science (e.g. nutrition, sensory and perception, eco-design, microbiology, biochemistry, process engineering) with data and knowledge in environmental analysis.

Besides the different domains of the available data, several other issues need to be addressed to be able to help experts in their decision process such as data heterogeneity, their different scale, their different purposes for which they have been collected and that can be complementary or even contradictory, their temporal evolution through the different unit operation. A major problem is the lack of cross domain studies and no clearly identified methodology about how to combine, aggregate and integrate data and knowledge in order to perform a multi criteria analysis in food production.

When assessing the quality of food product, it is important to have knowledge about product properties all along its production process, about different process parameters and the environmental impact of the whole production process. This section presents two use cases of data integration using ontologies in food production: @Web and PO .

3.1 @Web

@Web (Buche et al., 2013b) is a Web application designed to collect, integrate and quey experimental data extracted from scientific documents found on the Web. @Web implements a complete workflow to manage experimental data: extraction and semantic annotation of data from scientific documents, data source reliability assessment and uniform querying of the collected data stored in a database opened on the Web.

@Web relies on an Ontological and Terminological Resource (OTR) (Buche et al., 2013a) which guides scientific data semantic annotation and querying. An OTR associates a terminological component to an ontology in order to establish a clear distinction between the linguistic expressions in different languages (i.e. the term) and the notion it denotes (i.e. the concept) (Roche et al., 2009; McCrae et al., 2011; Cimiano et al., 2011). For instance, English terms “Ethylene vinyl alcohol” and ‘‘EVOH” and the French term “Ethylène alcool vinylique” denote the same symbolic concept Ethylene_vinyl_alcohol. The @Web OTR is designed to

model scientific experiments. It is composed of two layers: a generic one and a specific one dedicated to a given application domain. The @Web OTR allows n-ary relations relevant for a given application domain to be defined. Those n-ary relations are used to annotate data in tables. In (Buche et al., 2013a) the @Web OTR for the Food packaging domain is presented (see Figure 2 for an example in this domain). In (Lousteau-Cazalet et al., 2016) the generic part of the @Web OTR was reused and its specific part was defined for the biorefinery process.

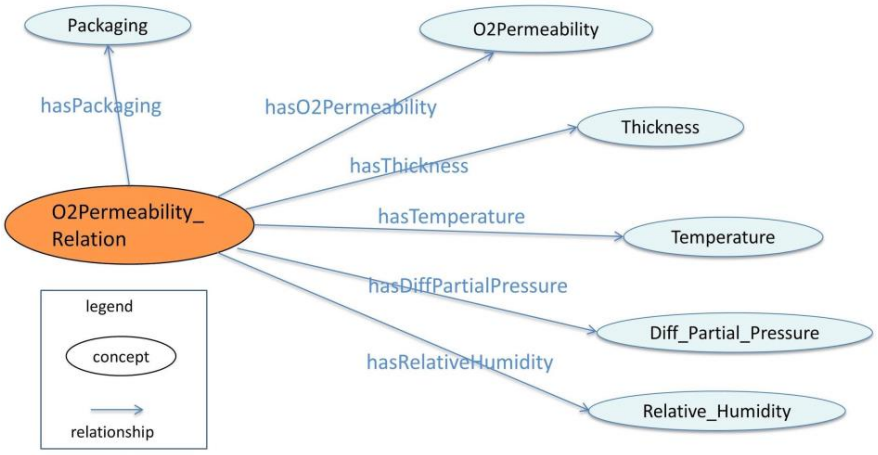


FIGURE 2 - *An example of ontological conceptual modelling: the O_2 permeability n-ary relation concept.*

@Web application is composed of two sub-systems. The first one is an annotation sub-system for the acquisition and annotation, with concepts of the OTR, of experimental data extracted from scientific documents; those annotated data are being stored into a database. This sub-system also allows the reliability of data sources to be assessed using the approach of (Destercke et al., 2013). The second sub-system is a flexible querying system based on the approach presented in (Destercke et al., 2011). @Web is implemented using the semantic web standards (RDF, OWL, SPARQL): the OTR is defined in OWL2-DL, annotated tables in XML/RDF and the querying in SPARQL. Section 4 will present a decision support system using the outcome (i.e. the database) of the @Web application.

3.2 PO²

PO² (Process and Observation Ontology) (Ibanescu et al., 2016) allows one to represent a food transformation process described by a set of experimental observations available at different scales and changing over time through the different unit operations of the production process. PO² ontology is the outcome of the data and knowledge representation tasks of projects involving INRA researchers. The goal of one of these projects, the CellExtraDry project, was to build a multicriteria decision support system allowing to improve the environmental impact for the production of stabilized micro-organisms (e.g. yeast). The goal of a second project, the NutriSensAl project, was to collect and integrate data from the cheese production system and from the domain concerning the sensorial perception of food; these NutriSensAl data will be used to implement a decision support system allowing to produce new foods with better nutritional qualities and acceptable by the consumers.

PO core ontology has been developed using the Scenario 6 of the NeON methodology (Suarez-Figueroa et al. 2012), i.e. reusing, merging and re-engineering ontological resources. PO core ontology has been built from SSN/SOSA, QUDT (http://qudt.org/schema/qudt/), OWL-TIME (http://www.w3.org/2006/time#) and @Web OTR described in Section 3.1. Moreover, and for the goal of increasing semantic interoperability, particularly in the life

sciences domain, PO core ontology was fully integrated with the Basic Formal Ontology (BFO), a small, and genuine upper level ontology.

PO core ontology concepts and relations are given in Figure 3. Three types of concepts are used:

- the concepts **Process**, **Itinerary** and **Step** concern the description of the production process. An itinerary is composed of a set of steps. A step describes an operation unit and it is characterized by its participants and its temporal duration.
- the concepts **Product**, **Mixture**, **Material** and **Method** concern the participants involved in a production step. A **Mixture** is the aggregation of several raw products. Material represents all the objects which are used during a step. These materials can be sensing devices performing measurements or transformation equipments. A method is the description of the way the observation has been performed.
- the concepts **Observation**, **Attribute** and **Scale**. An observation concerns an attribute, e.g. pH or temperature.

PO core ontology v2.0, implemented in OWL 2, is published on the AgroPortal ontology library (Jonquet et al. 2018) (http://agroportal.lirmm.fr/ontologies/PO2).

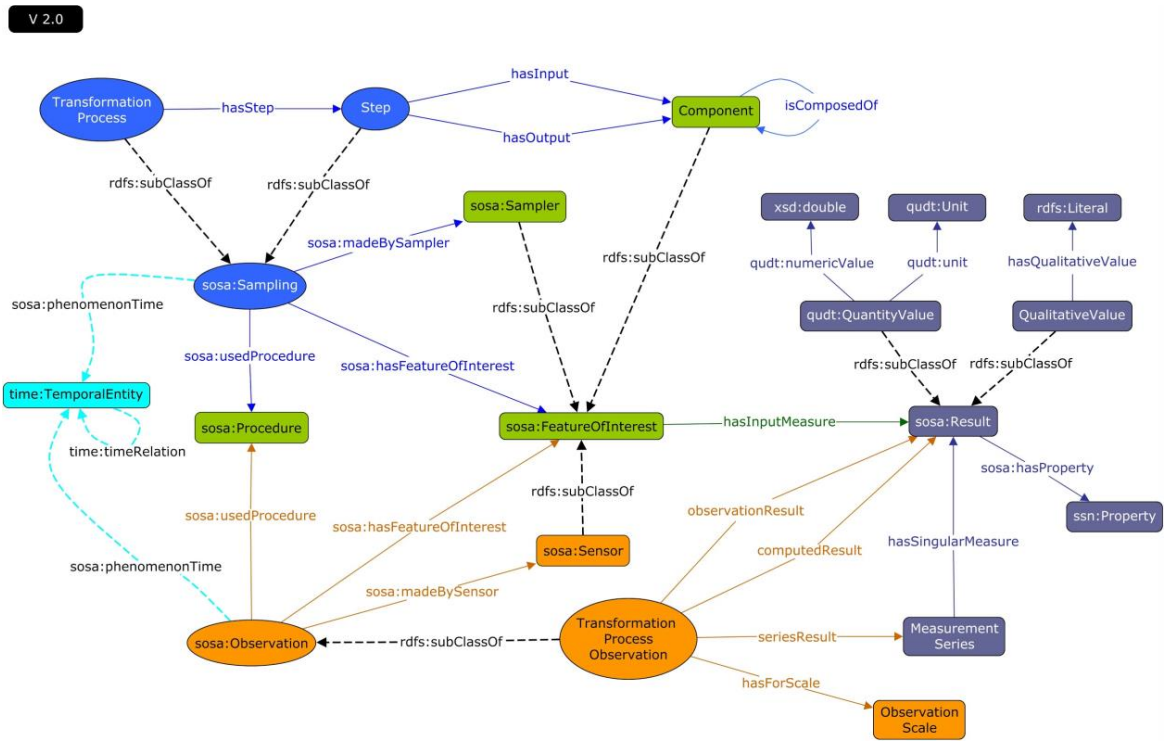


FIGURE 3 - *The core ontology PO².*

PO DG (DG for Dairy Gel) is a domain ontology built from PO core ontology in the field of dairy products (http://agroportal.lirmm.fr/ontologies/PO2-DG/). PO DG reuses concepts from Global Agricultural Concept Scheme (GACS) (http://agroportal.lirmm.fr/ontologies/GACS) which is aligned with AgroVoc (http://agroportal.lirmm.fr/ontologies/AGROVOC) and NALT (http://agroportal.lirmm.fr/ontologies/NALT). It allows to describe the production of cheese with attributes for sensory perception (e.g. texture or taste intensity) and rheological properties studied in the NutriSensAl project. Moreover PO DG ontology includes concepts needed to quantify the environmental impact using Life Cycle Assessment in the production of stabilized micro-organisms defined in the CellExtraDry Project.

PO DG contains 3475 concepts and 122 relations and it is available from the AgroPortal repository (http://agroportal.lirmm.fr/ontologies/PO2_DG) under the licence Creative Commons Attribution International 4.0 International (CC BY 4.0). Figure 4 gives an excerpt of the **Product** concept hierarchy in PO DG.

In (Penicaud et al. 2019), the presented results illustrate how the common vocabulary and the structure provided by the PO DG ontology allow the data combination from different domains and how this semantic approach can be useful to estimate missing data in NutriSensAl project. Moreover, an illustrative example is given to show how to transfer knowledge from the CellExtraDry project to the NutriSensAl project in order to evaluate the environmental impact using Life Cycle Assessment (LCA) by giving hints to the LCA expert about relevant parameters to be measured. A RDF repository `PO²DG_dataset` is under the process of data integration and will store the available data of the NutriSensAl project structured using the PO DG ontology. The decision support system build upon the integrated and semantically annotated data from the `PO²DG_dataset` is under construction.

The next section presents an example of a decision support system build upon the integrated and semantically annotated data presented in Section 3.1.

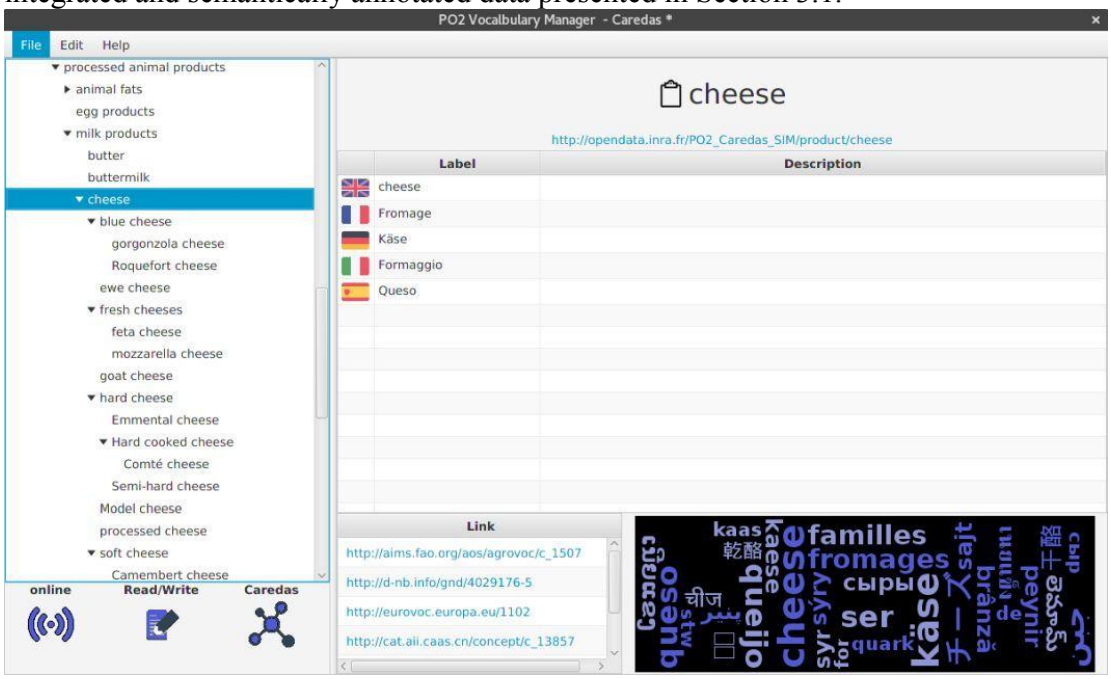


FIGURE 4 - *An excerpt of the **Product** concept hierarchy in PO²DG.*

4 Multicriteria decision support based on integrated data and food chain stakeholders' preferences and constraints

Knowledge engineering methods have been used to mimic the decision making process of the human brain and allow the development of decision support systems (DSS) like EcoBioCAP software that help users take the right decision in the field of food packaging (Guillard et al. 2015). EcoBioCAP is a powerful DSS tool able to answer a complex multi-criteria query such as: "I want a packaging material that will maintain the quality of strawberries (i.e. with the permeability properties that match the respiration of strawberries), at a cost of less than €3 per kg, and if possible transparent and derived from renewable resources". Flexible querying methodologies employed in knowledge engineering were used to develop this tool (Destercke et al. 2011). Below we briefly present the EcoBioCAP DSS tool that have

been developed at the junction between different fields of expertise such as food engineering, computer science, knowledge engineering, argumentation and numerical simulation.

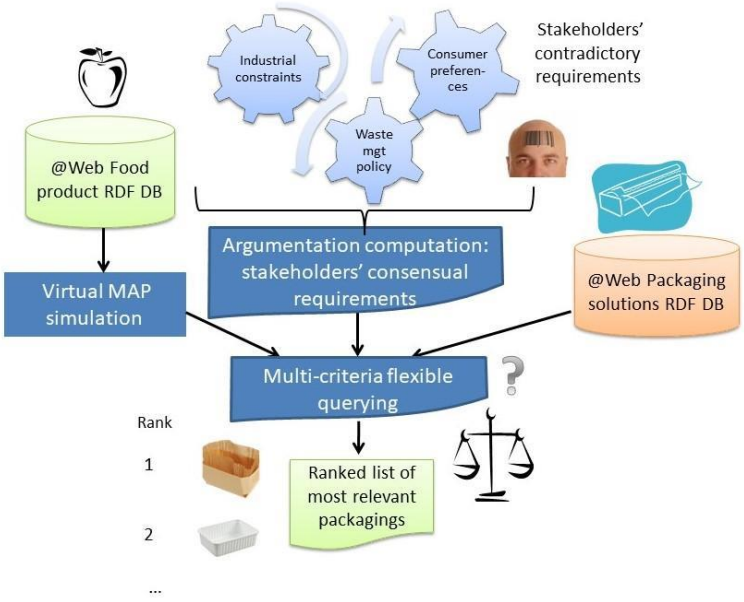


FIGURE 5 - *Architecture of EcoBioCAP DSS*.

EcoBioCAP DSS tool (see Figure 5) relies on data, the food product characteristics (O2/CO2 respiration) and the food packagings properties (O2/CO2 permeabilities), stored in the @Web RDF triplestore presented in Section 3.1. This tool retrieves respiration characteristics from the @Web food product database and uses this data plus other user-entered characteristics such as pack geometry to compute the optimal permeabilities for the food product.These permeabilities are automatically considered as mandatory preferences associated with selection criteria for the query, to which are added other mandatory or optional preferences that are determined by the user. The flexible querying module polls the @Web packaging database to retrieve the material that best satisfies the query preferences, and proposes as output a ranking of these materials. The DSS can manage both imprecise and missing data (Destercke et al. 2011). An answer is guaranteed even if no material satisfies the mandatory criteria. This type of tool marks a significant breakthrough, as it had never before been attempted in the field of food packaging. The first step in the process of building a DSS in the field of food packaging is to develop the numerical program that will serve to compute the evolution of food quality in relation to mass transfers in the food/packaging system. Several mathematical models have been developed that combine mass transfer models (based on Fick's laws) with food degradation models, such as the Mickaëlis-Menten equation for respiration or first-order reactions for oxidation (Penicaud et al. 2011). The EcoBioCap numerical model (Guillard et al. 2016) is used to adjust the packaging material to "the strict minimum", i.e. just those mass transfer properties necessary to maintain the protective atmosphere within a given range of values. Mathematical models for food engineering do allow some technical outputs to be computed but are not sufficient for decision making in an industrial world where choice of a packaging material is a multi-criteria decision. To take into account this aspect, the EcoBioCAP tool was developed to choose the most suitable packaging material for respiring produce from a dedicated database by answering bipolar multi criteria querying (currently four criteria considered in the first prototype). Bipolarity refers to the human reasoning that combines information on pros with information on cons to make decisions, choices or

judgments. Some preferences are modeled as constraints for which satisfaction is mandatory, while others are 'nice-to-haves' for which satisfaction is optional. Any packaging material that fails to satisfy the constraints is definitively discarded, while preference for a packaging increases the more it satisfies optional nice-to-haves. It is thus natural, in this context, for the querying process to make use of a bipolar approach, as it can handle compound preferences made of mandatory conditions and optional conditions. The web application and short demo videos are available at https://umr-iate.cirad.fr/equipes/ico/resultats-marquants/ecobiocap-dss.

We have seen that the EcoBioCap DSS tool is able to use preferences associated with multiple criteria in order to refine the decision. In the previous paragraph, we have mainly discuss technical criteria as optimal permeabilities values. Additional criteria have to be taken into account to obtain a sustainable decision. Those criteria are based on stakeholders' preferences (by example consumers' preferences about the packaging's transparency or packaging's end-of life) in the context of collective decision making, as different stakeholders might have different preferences. In such a situation, it is important to be able to input preferences that represent as accurately as possible the individual desiderata. To solve this, preferences aggregation techniques, i.e. voting rules, have been designed to merge a set of individual preferences into a unique, "global", preference. The Papow software (Karanikolas et al. 2018, Jedwabny et al. 2019) implements several of these voting rules, such as plurality, k-approval or Borda count, and allows for an in-depth exploration of individual preferences through filtering and clustering of the voters.

However, while aggregating preferences provides a direct answer to the collective decision-making problem, it only offers a partial solution with various shortcomings such as the inability to explain why agents' preferences differ in the first place. In order to understand these disagreements, it is then necessary to study the justifications behind the preferences, i.e. the arguments an individual might put forward to support her preferences.

The DAMN software (https://hamhec.github.io/damn/home) (Hecham et al. 2018) has been implement to meet this need. Indeed, by using logical reasoning techniques, this software is able to compute justified preferences through automatic justification analysis. The process is the following: first, the participants are asked to provide a justification for each of their preferences; then agreements and contradictions between the reasoning steps provided by these agents are automatically detected; finally, the participants can discuss the diverging reasoning steps and potentially change their preferences or justifications. As a result, DAMN provide collectively assessed preferences that can then be aggregated thanks to Papow in order to obtain a unique, collectively justified, preference which can be considered as input of a query which is executed by the EcoBioCap DSS tool presented in Figure 5.

5 International initiatives to share agri-food ontologies and other semantic resources

The ontologies presented in this paper are both research materials and results. As such, it is important to, at a minimum, make them public, or at best make them reusable. The reasons to open our ontologies are many, including:

- data reusability: semantic resources used to structure or document a dataset should be made explicit, be findable, and accessible.
- data interoperability: ontologies should be made public and actionable for the construction of concept mappings which contribute to the semantic interoperability of datasets or systems using them
- research reproducibility and process transparency: (successive versions of) ontologies used in decision support systems or reasoner should be accessible and actionable by anyone willing to run the system

- avoidance of duplicate effort: the expertise and time spent on the formalisation of domain knowledge should be shared with the secondary effect of increasing the quality and modularity of semantic resources.

Recently, a repository and a catalogue have been created to respectively host and reference the semantic resources that are useful to research and industry in the agricultural and nutrition domain. Their use has shown that such domain specific portals contribute to the structuring of the community, and facilitate cross-discipline collaborations. Their development have been initiated or reinforced by initiatives such as Research Data Alliance (RDA) (https://rd-alliance.org/) or Global Open Data for Agriculture and Nutrition (GODAN) (https://www.godan.info/), presented and discussed at the end of this section.

5. 1 AgroPortal and the Agrisemantics Map of Data Standards

Many vocabularies and ontologies are produced to represent and annotate agronomic data. For instances, the Plant Ontology, Crop Ontology, and more recently, the Agronomy Ontology, Food Ontology, or Process and Observation Ontology presented in Section 3.2.

Semantic interoperability is a key issue for agronomy, and the use of ontologies a way to address it (Lehmanna et al. 2012), Ontologies have opened the space to various types of semantic applications, to data integration (Buche et al. 2013b), to process and transformation description (Lousteau-Cazalet et al. 2016, Muljarto et al. 2017) or decision support (Guillard et al. 2015).

However, those ontologies are spread out over the web (or even unshared), in many different formats and types, of different size, with different structures and from overlapping domains. Therefore, there is need for a common platform to receive and host them, align them, and enabling their use in agro-informatics applications. There exists a need of a one-stop-shop for ontologies in the agri-food domain enabling to identify and select an ontology for a specific task as well as offering generic services to exploit them in search, annotation or other scientific data management processes. The need is also for a community-oriented platform that will enable ontology developers and users to meet and discuss their respective opinions and wishes.

The AgroPortal project, is a community effort started in 2015-2016 by the Montpellier scientific community (LIRMM, IRD, CIRAD, INRA, Bioversity International) to build an ontology repository for agronomy and related domains. Our goal is to facilitate the adoption of metadata and semantics to facilitate open science and the production of FAIR data. By enabling straightforward use of existing ontologies, we expect data managers and researchers to focus on their tasks, without requiring them to deal with the complex engineering work needed for ontology management.

Mid-2015, by reusing the NCBO BioPortal technology (Noy et al. 2009), we have designed AgroPortal (http://agroportal.lirmm.fr), an ontology repository for the agronomy domain but also food, plant, and biodiversity sciences. AgroPortal (Jonquet et al. 2018) offers a robust and reliable advanced prototype that features ontology hosting, search, versioning, visualization, comment, and recommendation; enables semantic annotation, stores and exploits ontology alignments, and enables interoperation with the semantic web. By specifically addressing the requirements of the agronomy community, AgroPortal has kindled an important interest both at the national and international levels. The platform currently hosts 106 vocabularies with more than 2/3 of them not present in any similar ontology repository and 10 private ontologies. We have identified 80 other candidate ontologies that will be loaded in the future to complement this valuable resource. The platform already has more than 100 registered users and some vocabularies are visited more than 100 times per month.

In addition to its core repository of ontology mission, AgroPortal also offers many applicable tools, including a mapping repository, an annotator, an ontology recommender, and

community support features. Our vision was to adopt, as the NCBO did, an open and generic approach where users can easily participate to the platform, upload content, and comment on others' content (ontologies, concepts, mappings, and projects).

However, the current AgroPortal prototype only partially addresses the needs of the community: it is not multilingual, it is limited in terms of ontology alignment capabilities and does not provide semantic-search and retrieval of ag & biodiv data. Addressing some of these issues are among the objectives of a ANR-funded project (starting in June 2019) called D2KAB (Data to Knowledge in Agronomy and Bioidiversity – www.d2kab.org).

In parallel, and based on previous work at Food and Agriculture Organization of the UN (UN-FAO), the Agrisemantics Map of Data Standards (http://vest.agrisemantics.org) (Pesce et al. 2018) is a catalog of semantic resources of interest for agri-food. This catalog lists any kind of data standards (not only ontologies & vocabularies) and we have developed semi-automatic synchronization mechanisms mainly to import ontologies and vocabularies metadata from AgroPortal to the Map.

The objectives of building the Agrisemantics Map of Data Standards task were: (i) To map currently available open and proprietary standards in use for the exchange of key data on agriculture and nutrition; (ii) To identify where a lack of standards is inhibiting the effective use of agricultural and nutritional data and the best methods for promoting open data standards. A gap analysis was also produced by GODAN Action.

In 2018, both AgroPortal and Agrisemantics Map of Data Standards content were incorporated in the FAIRsharing catalog (https://fairsharing.org/) for a larger dissemination.

5.2 Scientific and technical international fora

Exposing semantic resources in catalogues and repositories ensures their findability and accessibility by the community of semantic professionals of the agricultural domain and associated disciplines. But fostering their reuse in other contexts as well as reaching new users need to be supported by networking activities like scientific and technical meetings, and task force groups. The Research Data Alliance (RDA) and the Global Open Data for Agriculture and Nutrition (GODAN) initiative are two examples of international fora that have supported the social part of our work for more than five years. These are complementary to more classical scientific fora we are involved in like the IN-OVIVE International Workshop on sources and data integration in agriculture, food and environment using ontologies, or the AgroSem Track at the Metadata and Semantic Research (MTSR) conference.

For research organisations like ours, RDA represents the international arena where to check for latest advancements on the data side, gain inspiration for new ideas, and validate ongoing approaches. It is also the place to promote the approaches developed in our labs to varied stakeholders of the data chain (application developers, funders, data managers, etc.). Since the launch of RDA in 2013, the Interest Group on Agricultural Data (IGAD) has been a forum for sharing experience and providing visibility to research and work in agricultural data. With over 200 members, it represents stakeholders in managing data for agricultural research and innovation, including producing, aggregating and consuming data. Meetings every 6 months reinforce shared knowledge, technological transfer, and stir new collaborations while working groups created under its umbrella allow to develop a community approach to address data related issued.

In 2017, with a group of public and private organisations including INRA, Irstea, LIRMM, IRD, Embrapa, Wageningen University, UN-FAO, CABI, Agroknow and AgGateway, we launched the Agrisemantics Working Group to allow semantic experts and agriculture stakeholders meet and collaborate. There was the shared idea that results of scientific research and developed methodologies like those described above need to reach a

larger community of adopters. Indeed, while proved to efficiently address some complex data interoperability issues, semantic technologies globally suffer from 1) being perceived as too difficult, requiring advanced competencies, and 2) a lack of tools and methodologies allowing non-semantic experts to seamlessly edit, align, and use semantic resources on their data. A landscaping exercise (https://www.rd-alliance.org/system/files/documents/Deliverable1%20-%20Landscaping.pdf) and collection of requirements (Caracciolo et al. 2018, https://www.rd-alliance.org/deliverable-2-use-cases-and-requirements) by Agrisemantics substantiated this observation for the agricultural and nutrition community. Based on an open dialog with the community of researchers and practitioners of semantic technologies, we have identified a few basic points to address in order to make their use more widespread and effective. Then, we have translated those points in specific recommendations roles and activities in data stewardship (report under review by RDA secretariat for endorsement).

In the meantime, the GODAN initiative (https://www.godan.info/) announced at the Open Government Partnership Conference in October 2013, has contributed to the structuring of the community by supporting working groups (on soil data, nutrition data gap, capacity building, etc.), and specific actions like the Agrisemantics Map of Standards mentioned above. Through the promotion of success stories and the publication of reports and white papers, GODAN advocates for high level political and policy actions that enable action on the ground.

Since 2016, the FAIR principles (Wilkinson et al. 2016) have emerged as a supporting framework on the path of open science. While the FAIR principles are generic - aiming at making data (code, services, etc) more Findable, Accessible, Interoperable, and Reusable - their adoption and implementation require deep anchorage into the domain communities. To achieve this, on the initiative of INRA, Wageningen University and Research, and GODAN, has started a GO FAIR Implementation Network called "Food Systems" (https://www.go-fair.org/implementation-networks/overview/food-systems/). Our objective is to advance a global data ecosystem for agriculture and food by implementing FAIR data and services. In particular, building on the outputs of the RDA Agrisemantics Working Group and their knowledge and experience in the field, the French partners of the Food Systems IN will use the Food Systems IN to make their ontologies and other semantic resources more FAIR, as well as the data that use them. We will also collectively address the need for training on semantic technologies and methodologies with the aim of reaching data managers and application developers in particular.

6 Conclusion

We have shown in this article through different French computer science research works and initiatives some of the main locks to face and some proposed solutions using ontologies in data and knowledge integration for decision support in the Agri-food domain. Our aim was not to be exhaustive but to present illustrative examples in the Agri-food domain.

Several questions remain and we are only at the beginning of the data revolution in the Agri-food domain. First, networks of ontologies are promising answers (Muljarto et al. 2017) to link data from primary production to data from secondary production allowing therefore to be able to assess the impact of operations undertaken during primary production on the final quality of food product. Second, the data alignment problem remains a very challenging issues to face in particular in the Agri-food domain composed of many heterogeneous sub-domains that deal with very specific data treated at different scales by many distinct disciplines with their own objectives, scopes and methods. Third, the encounter of two worlds of data will have to be taken into account: the big data flow coming from more and more sophisticated equipment and the experimental data made by human in laboratories with many missing values. In this second kind of data, when trying to estimate missing data using available ones, important domain questions rise: i) what method can be used for the estimation? ii) what available data

should be used? and iii) what are "similar" data? Domain experts need to find answers to these questions and semantic techniques may help. The question *what are "similar" data?* is in particular very challenging in computer science and is investigated more and more in the context of linked data (Halpin et al. 2010; Beek et al. 2018).

References

BATTLE R. & KOLAS K. (2012). Enabling the geospatial Semantic Web with Parliament and GeoSPARQL. Semantic Web 3(4): 355-370.

BEEK W., RAAD J., WIELEMAKER J. & VAN HARMELEN F. (2018). sameAs.cc: The Closure of 500M owl: sameAs Statements. In ESWC 2018, p. 65-80.

BOWERS S. (2012). Scientific workflow, provenance, and data modeling challenges and approaches. In J. Data Semantics 1 (1), p. 19–30.

BUCHE P., DERVAUX S., DIBIE-BARTHELEMY J., SOLER L., IBANESCU L. & TOUHAMI T (2013a). Intégration de données hétérogènes et imprécises guidée par une ressource termino-ontologique. In Revue d'Intelligence Artificielle 27(4-5), p. 539-568.

BUCHE P., DIBIE-BARTHELEMY J., IBANESCU L. & SOLER L. (2013b). Fuzzy Web Data Tables Integration Guided by an Ontological and Terminological Resource. In IEEE Trans. Knowl. Data Eng. 25(4), p. 805-819.

CARACCIOLO C., AUBIN S.,WHITEHEAD B. & ZERVAS P. (2018). Semantics for Data in Agriculture: A Community-Based Wish List. In Contributions to Management Science (pp. 340–345). Springer International Publishing. https://doi.org/10.1007/978-3-030-14401-2_32.

CIMIANO P., BUITELAAR P., MCCRAE J. & SINTEK, M. (2011). LexInfo: a declarative model for the lexicon-ontology interface. In J. Web Sem. 9 (1), 29–51.

COMPTON M., BARNAGHI P., BERMUDEZ L., GARCIA-CASTRO R., CORCHO O., COX S., GRAYBEAL J., HAUSWIRTH M., HENSON C. & HERZOG A. (2012). The ssn ontology of the w3c semantic sensor network incubator group. In Web semantics : science, services and agents on the World Wide Web, 17, 25–32.

COX S. (2011) Observations and measurements-xml implementation. https://publications.csiro.au/rpr/pub?pid=csiro:EP115858

DESTERCKE S., BUCHE P. & GUILLARD V. (2011) A flexible bipolar querying approach with imprecise data and guaranteed results. In Fuzzy Sets Syst 169:51–64. doi: 10.1016/j.fss.2010.12.014.

DESTERCKE S., BUCHE P. & CHARNOMORDIC B. (2013). Evaluating data reliability: an evidential answer with application to a web-enabled data warehouse. In IEEE Trans. Knowl. Data Eng. 25 (1), 92–105.

DOAN A., HALEVY A. & IVES Z. (2012). Principles of Data Integration. Morgan Kaufmann 2012, ISBN 978-0-12-416044-6, pp. I-XVIII, 1-497.

ETSI TS 103 410-2 - v1.1.1 (2017). SmartM2M ; Smart Appliances Extension to SAREF ; Part2 : Environment Domain. Rapport interne, ETSI.

GARDNER S.P. (2005). Ontologies and semantic data integration. In Drug Discov. Today 10 (14), 1001–1007.

GUARINO N. , OBERLE D. & STAAB S. (2009). What Is an Ontology? In Handbook on Ontologies, p. 1-17.

GUILLARD V., BUCHE P. & DESTERCKE S (2015). A decision support system to design modified atmosphere packaging for fresh produce based on a bipolar flexible querying approach. In Comput Electron Agric 111:131–139. doi: 10.1016/j.compag.2014.12.010

GUILLARD V., COUVERT O., STAHL V.& BUCHE P. et al. (2016). Validation of a predictive model coupling gas transfer and microbial growth in fresh food packed under modified atmosphere. In Food Microbiol 58:43–55. doi: 10.1016/j.fm.2016.03.011.

HALLER et al. (2017). Semantic Sensor Network Ontology W3C Recommendation 19 October 2017 (Link errors corrected 08 December 2017) https://www.w3.org/TR/vocab-ssn/.

HALPIN H., HAYES P., MCCUSKER J., MCGUINNESS D. & THOMPSON H. (2010). When owl: sameAs Isn't the Same: An Analysis of Identity in Linked Data. In International Semantic Web Conference (1) 2010: 305-320.

HECHAM A., BISQUERT P. & CROITORU M. (2018). On a Flexible Representation for Defeasible Reasoning Variants. In AAMAS 2018: 1123-1131.

HOBBS J. & PAN F. (2004). An ontology of time for the semantic web. ACM Trans. Asian Lang. Inf. Process. 3(1): 66-85.

HODGSON R. (2016). Quantities, Units, Dimensions and Types (QUDT) Schema - Version 2.0 http://qudt.org/doc/2016/DOC_SCHEMA-QUDT-v2.0.html

IBANESCU L., DIBIE J., DERVAUX S., GUICHARD E. & RAAD J. (2016). PO^2 - A Process and Observation Ontology in Food Science. Application to Dairy Gels. In MTSR 2016: 155-165.

JEDWABNY M., BISQUERT P. & CROITORU M. (2019). Papow Aggregates Preferences and Orderings to select Winners. In AAMAS 2019, to appear.

JONQUET C., TOULET A., ARNAUD E., AUBIN S., DZALE YEUMO E., EMONET V., & GRAYBEAL J. (2018) AgroPortal: a vocabulary and ontology repository for agronomy. In Computers and Electronics in Agriculture 144 (2018): 126-143.

KARANIKOLAS N., BISQUERT P., BUCHE P., KAKLAMANIS C. & THOMOPOULOS R (2018). A Decision Support Tool for Agricultural Applications Based on Computational Social Choice and Argumentation. In IJAEIS 9(3): 54-73.

LEHMANNA, R.J., REICHEA, R., SCHIEFERA, G. (2012). Future internet and the agri-food sector: State-of-the-art in literature and research. In Computers and Electronics in Agriculture. 89, 158–174 (2012).

LEBO T., SAHOO S. & MCGUINNESS D. (2013). PROV-O: The PROV Ontology. W3C Recommendation. http://www.w3.org/TR/2013/REC-prov-o-20130430/

LOUSTEAU-CAZALET C., BARAKAT A., BELAUD J.P., BUCHE P., BUSSET G., CHARNOMORDIC B., DERVAUX S., DESTERCKE S., DIBIE J., SABLAYROLLES C.& VIALLE C. (2016). A decision support system for eco-efficient biorefinery process comparison using a semantic approach. In Computers and Electronics in Agriculture 127: 351-367.

MCCRAE J., SPOHR D. & CIMIANO P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. In: ESWC, vol. 1, pp. 245–259.

MADIN J., BOWERS S., SCHILDHAUER M., KRIVOV S., PENNINGTON D. & VILLA F. (2007). An ontology for describing and synthesizing ecological observation data. In Ecological Informatics, 2(3), 279–296.

MULJARTO A., SALMON J-M., CHARNOMORDIC B., BUCHE P., TIREAU A. & NEVEU P. (2107). A generic ontological network for Agri-food experiment integration - Application to viticulture and winemaking. In Computers and Electronics in Agriculture 140: 433-442.

NOY N., SHAH N., WHETZEL P., DAI B., DORF M., GRIFFITH N. & JONQUET C. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse." In Nucleic acids research 37, no. suppl_2, W170-W173.

PENICAUD C., BROYART B. & PEYRON S. (2011). Mechanistic model to couple oxygen transfer with ascorbic acid oxidation kinetics in model solid food. In J Food Eng 104:96–104. doi: 10.1016/j.jfoodeng.2010.11.033.

PÉNICAUD C., Ibanescu L., ALLARD T., FONSECA F., DERVAUX S., PERRET B., GUILLEMIN H., BUCHIN S., SALLES C., DIBIE J. & GUICHARD E. (2019). Relating transformation process, eco-design, composition and sensory quality in cheeses using PO^2 ontology. In International Dairy Journal, vol 92, p. 1-10.

PERROT N., DE VRIES H., LUTTON E., VAN MIL H.G.J., DONNER M., TONDA A., MARTIN S., ALVAREZ I., BOURGINE P., VAN DER LINDEN E. & AXELOS M. (2016). Some remarks on computational approaches towards sustainable complex agri-food systems. In Trends Food Sci. Technol. 48, 88–101. http:// www.sciencedirect.com/science/article/pii/S0924224415002186.

PESCE V., TENNISON J., MEY L., JONQUET C., TOULET A., AUBIN S., & ZERVAS P. (2018). A Map of Agri-food Data Standards. In F1000 Research, Technical Report 7-177, Global Open Data for Agriculture and Nutrition (GODAN), February 2018.

POVEDA-VILLALON M., NGUEN Q.-D., ROUSSEY C., CHANET J.-P. & DE VAULX C. (2018). Ontological requirement specification for smart irrigation systems: a SOSA/SSN and SAREF comparison. In Proceedings of the 9th International Semantic Sensor Networks Workshop SSN2018, Monterey, USA,

ROCHE C., CALBERG-CHALLOT M., DAMAS L. & ROUARD P. (2009). Ontoterminology – a new paradigm for terminology. In: KEOD, pp. 321–326.

ROUSSEY C. & ABDERRAHMANI GHORFI T. (2018). Annotation sémantique pour une interrogation experte des Bulletins de Santé du Végétal. Dans les Actes des 29e Journées Francophones d'Ingénierie des Connaissances IC 2018, adossée à la 11e Plate-forme Francophone d'Intelligence Artificielle, 2018, Nancy, p 37-52. https://hal.archives-ouvertes.fr/hal-01839545

SEEDAH D.P.K., SANKARAN B. & O'BRIEN W.J. (2015). Approach to classifying freight data elements across multiple data sources. In Transp. Res. Rec. 2529 (18), 56–65.

SUÁREZ-FIGUEROA M.C., GÓMEZ-PÉREZ A., FERNÁNDEZ-LÓPEZM. (2012). The NeOn Methodology for Ontology Engineering. In Ontology Engineering in a Networked World 2012, p. 9-34.

WILKINSON M. D., DUMONTIER M., AALBERSBERG Ij. J., APPLETON G., AXTON M., BAAK A. & MONS B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. In Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18.

WOLFERT J., VERDOUW C., VERLOOP C. & BEULENS A. (2010). Organizing information integration in agri-food – a method based on a service-oriented architecture and living lab approach. In Comput. Electron. Agric. 70 (2), 389–405.

DOREMUS: un graphe d'œuvres musicales interconnectées

Manel Achichi,[1], Pasquale Lisena,[2] Konstantin Todorov,[1] Raphaël Troncy,[2]
et Jean Delahousse[3]

[1] LIRMM, University of Montpellier, CNRS, France
{achichi, todorov}@lirmm.fr

[2] EURECOM, Sophia Antipolis, France
{pasquale.lisena, raphael.troncy}@eurecom.fr

[3] OUROUK, Paris, France
delahousse.jean@gmail.com

Mots-clés : Ontologie, humanités numériques, interconnexion des données, jeu de données.

Résumé de [Achichi *et al.* (2018)], publié à ISWC 2018.

Trois grandes institutions culturelles françaises, la Bibliothèque Nationale de France, Radio France et la Philharmonie de Paris, ont uni leurs efforts à ceux d'universitaires en informatique et en sciences sociales dans le cadre du projet DOREMUS [1]. L'objectif principal était de développer des méthodes et outils pour décrire sémantiquement leurs catalogues d'œuvres et d'interprétations de musique classique et traditionnelle et de les ouvrir à la communauté Web en suivant les principes des données liées.

Une contribution majeure du projet est le développement de l'ontologie DOREMUS [Choffé & Leresche (2016)] qui étend les modèles CIDOC-CRM [Doerr (2003)] et FRBRoo [Doerr *et al.* (2008)] décrivant l'information bibliographique, en l'adaptant au domaine de la musique, comblant ainsi un important écart de représentation. Ainsi, nous avons défini un certain nombre de nouvelles classes et propriétés permettant de décrire les aspects d'une œuvre liés à la musique, tels que la clé musicale, le genre, le tempo, l'instrumentation tout en respectant les principes de modélisation de FRBRoo. Plusieurs vocabulaires contrôlés définissant des concepts propres à la musique (comme les genres musicaux ou les clés) ont été recueillis ou élaborés, interconnectés et publiés en utilisant le langage SKOS.

La construction du graphe de connaissances DOREMUS implique la mise en place d'un workflow qui inclut la conversion des données originales en RDF en suivant l'ontologie DOREMUS et l'interconnexion de ces données. Les données des catalogues des trois institutions partenaires sont disponibles au format MARC [2] ou XML. Des outils pour la conversion des données en RDF selon le modèle DOREMUS ont donc été développés : ils sont génériques comme `marc2rdf` (valable pour tous les formats MARC) ou ad-hoc et utilisés pour une sérialisation XML spécifique. Ce processus aboutit à la construction de plusieurs graphes de connaissances sur les œuvres et événements musicaux (un pour chaque source de données). Ces graphes ont été interconnectés à l'aide de *Legato*, un outil de liaison de données capable de gérer l'hétérogénéité des méta-données musicales. *Legato* implémente un algorithme basé sur la sélection et le classement des clefs afin d'améliorer considérablement la précision au niveau des liens générés. À des fins d'évaluation, un jeu de données de référence a été créé manuellement par les experts des bibliothèques et communiqué à la communauté du Web Sémantique dans le cadre de l'initiative OAEI [3]. Le processus de fusion de données aboutit à la construction d'un graphe pivot de référence des œuvres musicales communes aux trois institutions.

Enfin, un moteur de recherche exploratoire nommé *Overture* [4] a été développé. L'application fait directement des requêtes au point d'accès SPARQL et fournit des informations issues du graphe de

1. http://www.doremus.org
2. https://www.loc.gov/marc/
3. http://oaei.ontologymatching.org/
4. http://overture.doremus.org/

connaissances dans une interface utilisateur web. *Overture* propose également un système de recommandation basé sur la similarité entre artistes et entre œuvres en s'appuyant sur des plongements calculés sur le graphe DOREMUS.

Le jeu de données DOREMUS complet ainsi que les vocabulaires contrôlés et quelques exemples de requêtes sont disponibles à `http://data.doremus.org`. Le tableau 1 récapitule le nombre d'entités pour les classes les plus représentatives et fournit des détails sur la présence d'informations spécifiques.

classe	BnF	Philharmonie	Itema3	total
Oeuvre	135,818	9,005	8,319	153,142
- avec l'instrumentation	123,219	4,621	0	127,840
- avec la clé musicale	19,645	1,973	0	21,618
- avec le genre musical	128,497	3,820	8,071	140,388
- avec le compositeur	133,371	7,741	8,231	149,343
- avec la date de composition	91,566	5,712	4,856	102,134
- avec le numéro de catalogue	20,796	2,908	0	23,704
- avec le numéro d'ordre	11,598	1,612	0	13,210
- avec le numéro d'opus	21,836	1,985	0	23,821
Interprétation	15,784	784	1,531	18,099
- avec plus d'une oeuvre interprétée	0	713	1,277	1,990
Piste	0	6,538	18,273	24,811

TABLE *1 – Graphe de connaissances DOREMUS*

Remerciements. Ce travail a été partiellement financé par l'Agence Nationale de la Recherche (ANR) dans le cadre du projet DOREMUS (ANR-14-CE24-0020).

Références

ACHICHI M., LISENA P., TODOROV K., TRONCY R. & DELAHOUSSE J. (2018). DOREMUS : A Graph of Linked Musical Works. In *17th International Semantic Web Conference (ISWC)*, Monterey, USA.

CHOFFÉ P. & LERESCHE F. (2016). DOREMUS : Connecting Sources, Enriching Catalogues and User Experience. In *24th IFLA World Library and Information Congress*, Columbus, OH, USA.

DOERR M. (2003). The CIDOC Conceptual Reference Module : An Ontological Approach to Semantic Interoperability of Metadata. *AI magazine*, **24**(3), 75–75.

DOERR M., BEKIARI C. & LEBOEUF P. (2008). FRBRoo : a conceptual model for performing arts. In *CIDOC Annual Conference*, p. 6–18.

Interactive Interpretation of Serial Episodes: Experiments in musical analysis

Béatrice Fuchs[1], Amélie Cordier[2]

[1] Université de Lyon, UJML3, IAE, LIRIS, F-69 008, Lyon, France. beatrice.fuchs@liris.cnrs.fr

[2] Université de Lyon, Hoomano, LIRIS, F-69 100, Villeurbanne, France. amelie.cordier@hoomano.com

This is a summary of a paper presented at EKAW-2018 (Fuchs & Cordier, 2018).

The context of this work is the study of sequential data that can be represented with sequences of timestamped events. The aim is to explore these sequences with sequence mining to discover *serial episodes* which are frequent event subsequences that occur frequently in data (Mannila *et al.*, 1997). The domain of melodic analysis is studied in this work : the aim is to highlight the structure of a musical piece by discovering its main melodic patterns. The episodes produced by the miner are examined by a user generally an expert of the domain who have to identify relevant episodes and interpret them. Meanwhile in the interpretation step, the user has to face to a recurrent overabundance of mining's results which makes difficult the identification of interesting ones. There is a real need to adopt a rigorous approach to methodically manage this step and assist the user's work. For this, we propose a visual and interactive approach to assist the interpretation of serial episodes.

An Interactive approach to the interpretation of serial episodes

We propose to assist the interpretation task by managing combinatorial redundancy in order to focus on relevant episodes. The assistance combines iteratively ranking and filtering useless episodes to help focusing on relevant ones. It has been exemplified in the Transmute prototype, a web-based application enabling user's interaction with events sequences and serial episodes that are represented graphically on a timeline with customisable icons.

The interpretation process consists in the main iterative steps : ranking, selection and filtering. The user can choose measures to rank episodes and then select among them to display their occurrences in the sequence. When a choice is made, a filtering process is triggered to clean up other episodes that can no longer be selected following the previous selections of the user. Finally, the user can interpret the episodes by attaching them annotations and record the model resulting from the interpretation into a knowledge base.

The ranking of episodes is performed thanks to several objective interestingness measures which estimate the relative importance and compactness of the episodes in the sequence. The first measure is the event coverage indicator which is the number of distinct events of the occurrences of an episode. The second measure is the spreading indicator which is the number of events of the sequence in the time intervals of the episode occurrences. The noise indicator is the difference between these two previous indicators and corresponds to the number of events of the sequence in the time intervals of the episode occurrences. Temporal measures may also be used when event duration are known.

The selection of an episode by the user triggers the filtering process which is based on the event coverage of the selected episode. The remaining episodes are examined and occurrences having at least an event in common with the event coverage are discarded. The support is consequently updated and episodes whose support becomes less than the given frequency threshold are discarded. This results in removing combinatorial redundancy around the chosen episode and leads to a gradual diminution of the remaining episodes, allowing to the user a better focus on other episodes.

Experiments

The experiment aims to verify the ability of the approach to improve the ranking of episodes and as a consequence, to a lower effort from the expert. For this, three musical pieces have been chosen and for each of them, an expert gave the relevant episodes to find, which we name the expert episodes. The miner was launched with parameters to ensure the presence of the expert episodes and the mined episodes were ranked using interestingness measures. The ranks of the expert episodes in the mining results are used to assess the hypothetical effort of the expert. The smaller the rank, the less important is the effort of the user to find the expert episodes. Two situations are compared in two tables : without filtering and with filtering. Without filtering, the effort is the highest rank of expert episodes. With filtering, the effort is the sum of the lowest ranks of expert episodes, since the examination is resumed at the beginning after each filtering and ranking, each time an expert episode in found. Experiments show an important diminution ($> 80\%$) of the effort for the three pieces with the ranking-filtering strategy. A counterpart is that some expert episodes may be discarded, and the recall measure may be quite bad in some cases.

Discussion, limits

The combination of both filtering and ranking is conclusive but may sometimes lead to a lower recall. A first experiment has been conducted to test the usability of the Transmute prototype by users, but a full scale experiment with more complex pieces and domain experts remains to be conducted. This approach is suitable only if data lend to compression. Moreover, the Transmute prototype can not handle more than several thousands of results.

Related works

The sur-abundance of results in data mining is a major issue for a long time. Among related works we can mention the *compactness* that measures gaps in episode occurrences (Tatti, 2014), pattern selection based on their ability to compress data (MDL) : (Rissanen, 1978; Vreeken *et al.*, 2011) for itemsets and (Lam *et al.*, 2014) for sequences and finally *human in the loop* approaches (Bertini & Lalanne, 2009). More recent approaches claim that interestingness is subjective in essence and take into account the goals of the user and its knowledge (van Leeuwen, 2014) but out approach is quite different in the sense that the user has a more active role in the process.

Références

BERTINI E. & LALANNE D. (2009). Surveying the complementary role of automatic data analysis and visualization in knowledge discovery. In *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery : Integrating Automated Analysis with Interactive Exploration*, p. 12–20 : ACM.

FUCHS B. & CORDIER A. (2018). Interactive interpretation of serial episodes : experiments in musical analysis. In C. FARON-ZUCKER & C. GHIDINI, Eds., *Knowledge Engineering and Knowledge Management, 21st International Conference - EKAW-2018*, LNAI 11 313, p. 131–146, Nancy, France : Springer.

LAM H. T., MÖRCHEN F., FRADKIN D. & CALDERS T. (2014). Mining compressing sequential patterns. *Statistical Analysis and Data Mining*, **7**(1), 34–52.

MANNILA H., TOIVONEN H. & VERKAMO A. I. (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, **1**, 259–289.

RISSANEN J. (1978). Modeling by shortest data description. *Automatica*, **14**(5), 465 – 471.

TATTI N. (2014). Discovering episodes with compact minimal windows. *Data Min. Knowl. Discov.*, **28**(4), 1046–1077.

VAN LEEUWEN M. (2014). Interactive data exploration using pattern mining. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, p. 169–182. Springer.

VREEKEN J., LEEUWEN M. & SIEBES A. (2011). Krimp : mining itemsets that compress. *Data Mining and Knowledge Discovery*, **23**(1), 169–214.

Apports des ontologies aux systèmes de recommandation : état de l'art et perspectives

Yu Du[1], Sylvie Ranwez[1], Nicolas Sutton-Charani[1], Vincent Ranwez[2]

[1] LGI2P, IMT Mines Ales, Univ Montpellier, Ales, France
prenom.nom@mines-ales.fr

[2] AGAP, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France
vincent.ranwez@supagro.fr

Résumé : La recommandation consiste à proposer à un utilisateur un *produit* qui corresponde au mieux à son profil ou ses attentes, que celles-ci aient été exprimées explicitement ou déduites à partir d'interactions avec le système. La représentation de connaissance du domaine considéré sous forme d'ontologies, ou plus largement de bases de connaissance, peut constituer un réel atout pour améliorer les performances d'un système de recommandation automatisé. Pourtant, leur utilisation dans ce contexte reste marginale.

Cet article a pour but de dresser un état de l'art de l'utilisation des ontologies et des bases de connaissance dans le contexte des systèmes de recommandation. Il détaille les différentes phases du processus de recommandation qui peuvent bénéficier de l'apport des approches sémantiques et montre comment différents domaines d'application ont su en bénéficier. Il se termine par des pistes de recherche qui ouvrent la voie à de nombreuses innovations en matière de recommandation automatique.

Mots-clés : Ontologie, Bases de connaissance, Système de recommandation, Prédiction des préférences, Mesures sémantiques

1 Introduction

Les ontologies, structures formelles pour la manipulation des connaissances liées à un domaine particulier, et plus largement les bases de connaissance, peuvent constituer un support majeur de l'automatisation d'un processus de recommandation. En effet, qu'elles permettent de formaliser un profil, d'assurer l'interopérabilité des données, le partage des connaissances, ou l'émergence de nouvelles connaissances par inférence, elles peuvent intervenir à différentes étapes d'un processus de recommandation. Leur incorporation dans des systèmes de recommandation a ainsi apporté une réelle plus-value dans différents domaines tels que le tourisme (Benouaret, 2017; Al-Hassan *et al.*, 2011; Moreno *et al.*, 2013), le domaine éducatif (Obeid *et al.*, 2018; Tarus *et al.*, 2017), le domaine cinématographique (Carrer-Neto *et al.*, 2012), ou la santé (El-Dosuky *et al.*, 2012; Ali *et al.*, 2018). Principalement dans des applications basées sur l'analyse du contenu, elles permettent de caractériser les *items* [1] à recommander ou les centres d'intérêt des utilisateurs afin de proposer les items les plus pertinents pour un utilisateur donné. De façon plus générale, dans le cadre d'un système de Recherche d'Information (RI), la capacité à représenter formellement des terminologies (concepts) et leurs relations via une ontologie permet de combler le fossé sémantique entre les données brutes (*e.g.* un corpus) et de filtrer les informations et améliorer ainsi la qualité de la recherche.

L'attrait pour les systèmes de recommandation (notés SdR par la suite) n'a fait que croître ces dernières années, au rythme de l'augmentation de la quantité des informations numériques accessibles et de leur développement exponentiel, et il traduit désormais des enjeux industriels majeurs. De par leur nombre, il devient en effet plus difficile de filtrer ces informations pour trouver celles qui conviennent pour un besoin particulier, dans un contexte donné. La solution adoptée par des grandes entreprises comme *Amazon, Netflix, Youtube (Google), etc.*,

1. On emploi ici le terme général d'item, parfois aussi objet, communément admis dans la communauté pour désigner tout à la fois un article de vente, un produit, une musique, une vidéo ou un contenu textuel à recommander, par exemple.

est de solliciter, au sein d'un système englobant, plusieurs types de systèmes de recommandation qui interagissent pour fournir un résultat. Étant donné un utilisateur, un tel système possède la capacité de prédire ses goûts en termes d'items présents dans le catalogue, et ainsi de lui présenter de manière ordonnée, en guise de recommandation, les N items qui semblent lui correspondre le mieux (Schafer *et al.*, 1999). Parmi les approches des SdR existantes dans la littérature, la plupart adopte des techniques liées à l'apprentissage automatique (*e.g.*, classification, régression, *etc.*) pour effectuer leurs prédictions. Très peu reposent sur les ontologies.

Pourtant, dans un cadre plus général, la Recherche d'Information (RI), qui peut être considérée comme une des tâches liées à la recommandation, intègre depuis de nombreuses années une dimension conceptuelle. En effet, la recherche d'information conceptuelle interrogeait dès les années 80 (Schank *et al.*, 1981) et trouve des applications pratiques intégrant des ontologies depuis une dizaine d'années (Dragoni *et al.*, 2012). Ce type de recherche se base sur des connaissances expertes liées au domaine pour à la fois représenter les ressources et les requêtes des utilisateurs afin de filtrer, après une phase d'appariement, les ressources pertinentes pour un utilisateur donné. Les ontologies interviennent naturellement dans ce contexte puisqu'elles sont particulièrement bien adaptées pour modéliser et manipuler des connaissances d'experts, de par la structuration des concepts au sein d'une hiérarchie d'une part et les relations sémantiques entre ces concepts d'autre part (Sy *et al.*, 2012). Or dans certains domaines, la recommandation exige, elle aussi, la prise en compte de connaissances expertes. C'est le cas, par exemple, dans le domaine de la santé où la recommandation pourra tenir compte à la fois de cette connaissance experte et de données contextuelles sur le patient, collectées grâce à certains objets connectés. Le système de recommandation peut également être doté d'un moteur d'inférence qui permet de vérifier la consistance des ontologies et de les utiliser comme base de raisonnement dans le processus de recommandation.

Dans ce papier, nous nous focalisons sur les apports des ontologies, et plus généralement des modèles de connaissance, aux systèmes de recommandation. A notre connaissances, il n'existe pas d'état de l'art concernant l'apport des ontologies à un système de recommandation, hormis (Tarus *et al.*, 2018) qui se restreint au domaine du *e-learning*. L'état de l'art que nous proposons se veut générique, et détaille les différentes contributions apportées par des ontologies aux systèmes de recommandation. Dans la section 2, nous rappellerons et illustrerons les notions de base des systèmes de recommandation puis la définition des ontologies ainsi que leur usage dans des contextes liés à la recommandation. La section 3 détaillera les principaux usages des ontologies au sein d'un système de recommandation, en particulier pour la construction de modèles pour représenter les utilisateurs ou les items et la section 4 détaillera leur utilisation pour le calcul de mesures sémantiques. La section 5 ouvrira des pistes de recherche possibles dans le contexte de la recommandation basée sur une modélisation de connaissance. Enfin la section 6 terminera ce papier en dressant les conclusions.

2 Systèmes de recommandation, ontologies de domaine et bases de connaissance

Avant d'aller plus loin dans la description des interconnexions entre modèles de connaissance et systèmes de recommandation, nous rappelons les bases de chacun des deux domaines.

2.1 Système de recommandation

Si pour réaliser l'analyse et le filtrage de nombreuses données numériques il existe des traitements déjà anciens, liés notamment à la recherche d'information (*e.g.* moteur de recherche, base de données), les systèmes de recommandation (SdR) sont eux relativement récents. On observe une augmentation remarquable de l'intérêt qui leur est porté depuis une dizaine d'années (Ricci *et al.*, 2015), souligné par exemple par le succès du *Netflix Prize* (Bennett & Lanning, 2007), une compétition populaire du domaine. Typiquement, un système de recommandation repose sur trois éléments centraux : les *utilisateurs*, *i.e.* un ensemble d'individus avec des goûts différents qui interagissent avec le système ; les items (ou les *objets*) à

recommander, rassemblés dans un *catalogue* ; les *notes* (ou les *évaluations*) retournées par les *utilisateurs* exprimant leurs préférences pour certains items. Ces notes sont obtenues soit de façon explicite en demandant aux *utilisateurs* d'évaluer des items, soit implicitement (Oard & Kim, 1998) en tenant compte des actions ou des comportements des *utilisateurs* pendant leurs sessions d'utilisation du système (*e.g.* temps de visualisation d'une vidéo, de lecture d'un texte, sauvegarde d'une ressource ou achat d'un produit). Pour les notations explicites, deux types d'échelles sont répandues sur Internet : i) l'une utilise des notes de 1 à 5, cette forme de notation est favorisée par les sites de e-commerce tels que Amazon ou TripAdvisor, par exemple ; ii) l'autre, basée sur une notation binaire du type *j'aime, je n'aime pas*, est favorisée par l'usage sur les sites de réseaux sociaux (*e.g.* Facebook, Youtube).

En général, un système de recommandation procède en deux phases majeures et consécutives : i) la prédiction et ii) la recommandation. De ce fait, considérant un *utilisateur*, un tel système est tenu d'estimer et de prévoir ses préférences pour certains items avant de procéder à la recommandation. La prédiction des préférences des utilisateurs demeure cruciale pour une recommandation de bonne qualité. Nous catégorisons les SdR selon la technique qu'ils adoptent pour réaliser la phase de prédiction. Parmi celles qui ont été mentionnées dans (Ricci *et al.*, 2015; Lu *et al.*, 2015), nous nous focalisons sur les trois techniques de filtrage les plus utilisées, et proposons sur cette base trois grandes catégories de systèmes de recommandation : SdR dont le filtrage est basé sur le Contenu (notés *CBF pour Content-Based Filtering*), SdR de filtrage collaboratif (notés *CF pour Collaboratif Filtering*) et SdR de filtrage hybride.

Un système de recommandation doté d'une approche *CBF* repose sur la supposition qu'un utilisateur aimera des items qui sont similaires à ceux qu'il a aimés précédemment (Lops *et al.*, 2011). Dans cette approche, la première étape clé est de construire une description (profil) de chaque item. Chacun est ainsi représenté par un ensemble de caractéristiques descriptives. Par exemple pour un film on peut spécifier le *type* auquel il appartient (*comédie, action, historique*), ses *acteurs* ou encore son *réalisateur*. Étant donné un utilisateur, l'approche analyse d'abord l'ensemble des items qu'il a précédemment notés afin de constituer un profil pour cet utilisateur. Pour cela le système peut mémoriser les caractéristiques qui correspondent aux items qui ont obtenu les meilleures notes de cet utilisateur. En comparant ensuite ce profil utilisateur et ceux des items qu'il n'a pas encore évalués, le système est capable de lui proposer une sélection d'items qu'il est susceptible d'apprécier de par la similitude de leurs profils. Un inconvénient majeur des approches basées sur le contenu est la nécessité de déterminer les caractéristiques qui sont pertinentes en fonction du contexte et du domaine considéré. En effet, cette sélection peut parfois être multi-domaines : *e.g.* pour un produit touristique il faut tenir compte du produit mais également de son environnement (géographie, météo, contexte politico-social...) et il est difficile de cerner l'ensemble des critères utiles. Par ailleurs, pour certains items, la caractérisation repose sur des critères très subjectifs et donc difficiles à appréhender : *e.g.* dans le domaine de la musique, la description de nos morceaux préférés repose parfois sur des sensations difficiles à expliciter. De plus, le fait de recommander toujours les items similaires au profil d'utilisateur, *i.e.* ceux possédant les mêmes caractéristiques, l'empêche de recevoir des recommandations surprenantes (problème de la *sérendipité*). L'utilisateur peut avoir la sensation qu'on lui propose toujours la même chose, créant ainsi une lassitude qui diminue sa satisfaction (Iaquinta *et al.*, 2008). Pour pallier ce problème de sur-spécialisation, des techniques favorisant la diversité des recommandations doivent être intégrées au SdR. Ce point sera discuté dans la section 5.

Le filtrage collaboratif peut, dans une certaine mesure, résoudre ce problème. L'approche de *CF* repose sur l'hypothèse que les utilisateurs possédant des préférences similaires notent les items de la même façon (Schafer *et al.*, 2007). Contrairement à l'approche *CBF* qui utilise les caractéristiques des items pour former les profils, le filtrage collaboratif ne prend en compte que les *notes* associées à ces items. Typiquement, comme chaque utilisateur ne note qu'une partie des items, ces *notes* engendrent une matrice incomplète, appelée la matrice d'*utilisateurs-items*, dans laquelle chaque ligne (resp. colonne) représente un utilisateur (resp. item). Par conséquent, le problème de la prédiction revient à compléter cette matrice en inférant les données manquantes. Afin de prédire la *note* qu'un utilisateur pourrait donner à un item, le système détermine d'abord un ensemble d'utilisateurs similaires, ses *voisins*. La similarité entre utilisateurs est calculée en mesurant la similarité de leurs vecteurs de notation

(*e.g.* avec une similarité de type cosinus ou corrélation de Pearson). Pour effectuer sa prédiction, le système ne considère alors que les notes fournies par les utilisateurs voisins (Sarwar *et al.*, 2001). De ce fait, cette technique recommande des items qui ont été évalués positivement par les k utilisateurs ayant le système de notation le plus similaire à celui de l'utilisateur focal. Dans la plupart des cas, l'approche *CF* est plus performante que l'approche *CBF* en terme de précision des prédictions. Pourtant, elle souffre de certaines limites parmi lesquelles le démarrage à froid (*cold start*) qui reste un des défis importants : il s'agit de l'impossibilité pour un SdR *CF* de recommander un nouvel item (puisque personne ne l'a noté) ou de faire des recommandations pour un nouvel utilisateur (puisqu'il n'a noté aucun item, le SdR ne sait pas quels sont ses voisins). Dans ce cas, la technique de *CBF* semble plus robuste que celle de *CF* puisqu'elle prend en compte le contenu des items et pas uniquement leurs notes (Aggarwal, 2016).

Visant à bénéficier des avantages des différentes approches et à pallier leurs inconvénients, et donc améliorer la qualité de la recommandation, une approche hybride est souvent employée. Dans la majorité des cas, un système de recommandation hybride combine la technique de *CBF* avec celle de *CF* de façon à s'adapter au mieux au contexte de la recommandation. Si plusieurs méthodes ont été proposées pour réaliser cette hybridation, aucune ne prédomine actuellement. Le lecteur pourra se référer à (Burke, 2002) pour plus de détails.

Il existe également des adaptations qui ont été proposées dans la littérature pour faire face à une problématique spécifique, *e.g.* l'approche sensible au contexte, l'approche basée sur une session, l'approche sensible à la confiance, *etc.* Nous notons parmi ces méthodes, l'approche *KB (Knowledge-based)* (Burke, 2000), nommée également *approche conversationnelle* (Lemdani, 2016; Alec, 2016) qui consiste à interagir avec l'utilisateur en lui posant des questions sur ce qu'il cherche. En fonction de ses réponses et d'un modèle de connaissances, les questions suivantes sont adaptées. Une approche hybride peut donc être construite en combinant ces techniques avec les approches typiques (*CBF* et *CF*).

2.2 Ontologie

Dans le domaine de l'informatique et plus précisément de l'ingénierie des connaissances et de l'intelligence artificielle, une ontologie peut être définie comme une spécification formelle et explicite d'une conceptualisation partagée (Gruber, 1993; Studer *et al.*, 1998). *Conceptualisation* fait référence à un modèle abstrait d'un domaine donné qui identifie les concepts représentatifs de ce domaine. *Explicite* signifie que le type de concepts utilisés et les contraintes sur leur utilisation doivent être explicitement définis. *Formelle* se réfère au fait qu'une ontologie doit être compréhensible par la machine, *i.e.*, cette dernière doit être capable d'interpréter la sémantique de l'information fournie. *Partagée* indique que l'ontologie supporte la connaissance consensuelle, et qu'elle n'est pas restreinte à quelques individus mais acceptée par un large groupe (Broekstra *et al.*, 2002). Le modèle ainsi obtenu constitue un graphe de connaissance du domaine dont les nœuds représentent des concepts et les arcs les relations. On parle de *T-Box* ou *Terminological Box*.

On peut également peupler cette ontologie en associant aux concepts des instances réelles qui les illustrent. On rajoute pour ce faire des assertions qui traduisent des *faits* du monde réel conformes au modèle conceptuel défini dans la *T-Box*. On parle alors de *A-Box* ou d'Assertion Component. L'ensemble (*T-Box* et *A-Box*) constitue une base de connaissance.

La construction d'une base de connaissance débute donc par la définition de l'ontologie. Pour ce faire il est nécessaire de définir un ensemble de concepts représentatifs du domaine qu'elle représente. Par exemple, pour le domaine cinématographique, on emploie souvent des concepts comme *Film, Acteur/Actrice, Réalisateur, Genre, etc.* Ensuite, des relations doivent être établies entre ces concepts. On distingue deux types de relation : les relations *taxinomiques* et les relations *sémantiques*. Le premier type structure la hiérarchie de concepts de l'ontologie en établissant des liens de spécificité ou généricité entre-eux, *e.g. Film d'action* et *Film*. Le second type représente une relation sémantique entre deux concepts, *e.g.* la relation *est réalisé par* entre *Film* et *Réalisateur*. Ensuite, une phase de peuplement permet d'associer les instances à ces concepts, *e.g. Titanic, Matrix, Spider-man* peuvent être rattachés au concept *Film*.

Pour rendre effective cette modélisation d'un domaine, *i.e.* l'ontologie, des langages de représentation de connaissances tels que RDFS et OWL sont utilisés. Ils possèdent tous un mécanisme d'inférence ou de raisonnement permettant de déduire des connaissances supplémentaires entre des éléments (*e.g.* instances, concepts) définis dans l'ontologie. De ce fait, l'ontologie permet d'inférer de nouvelles connaissances, ce qui est utilisé dans de nombreuses applications liées, entre autres, à l'aide à la décision.

Nous avons évoqué plus haut les agents conversationnels qui peuvent être utilisés pour affiner le profil des utilisateurs à partir de questionnaires adaptés en fonction d'un modèle de connaissance. Mais les modèles sémantiques peuvent également être exploités à d'autres phases du processus de recommandation. Ainsi (Ali *et al.*, 2018) proposent un système qui recommande des aliments et des médicaments aux patients atteints de diabète dans le contexte d'un accompagnement médical basé sur l'Internet des objets (*Internet of Things : IoT*). Leur approche exploite deux ontologies, *i.e.* l'ontologie des patients et l'ontologie des capteurs, grâce à une méthode issue de la logique floue afin d'inférer l'état d'un patient (sa condition à un instant donné) et ainsi lui faire des recommandations alimentaires adaptées. Le système proposé par (Chen *et al.*, 2012) utilise quant à lui des ontologies et un langage de règles (SWRL, *Semantic Web Rule Language*) afin d'assister les médecins dans leurs prescriptions et ainsi proposer aux patients les médicaments anti-diabétiques les plus appropriés.

Dans ce qui suit, nous ne considérons pas les systèmes pour lesquels la recommandation est uniquement constituée des résultats d'une recherche d'information, fut-elle conceptuelle. Ce champs de recherche, largement développé dans la littérature est en effet à la marge de notre champ d'étude qui se focalise sur les systèmes de recommandation qui reposent sur le triptyque *utilisateurs-items-notes*. Dans la section suivante, nous détaillons, dans ce contexte, les apports des ontologies et plus largement des bases de connaissance, aux systèmes de recommandation. L'annexe B présente une vue globale du processus de recommandation et indique les différents éléments du système qui peuvent tirer bénéfice des modèles sémantiques.

3 Annotation conceptuelle pour représenter les utilisateurs ou les items

L'apprentissage des profils des utilisateurs demeure une étape cruciale pour les systèmes de recommandation *CBF* et Hybrides. La façon dont ils sont construits est donc déterminante dans les performances du système. Dans une approche typique de *CBF*, un profil d'utilisateur (resp. d'item) est représenté par un vecteur dont la taille est égale au nombre de caractéristiques considérées et dont la valeur à l'indice i reflète l'importance de cette caractéristique pour cet utilisateur (resp. item). Pour créer ces profils, on s'appuie souvent sur des techniques provenant du *Traitement Automatique du Langage Naturel* (TALN), *e.g.*, le pré-traitement du corpus, l'extraction des termes représentatifs, la pondération des termes en fonction de leur fréquence relative (*e.g.*, TF-IDF). De ce fait, la méthode *CBF* est naturellement adaptée pour recommander des objets textuels (*e.g.* livres, articles, pages Web). Cette manière vectorielle de représenter les profils, qui se base sur des mots clés, possède un inconvénient majeur : elle ne capture pas la sémantique du profil étant principalement dictée par des opérations d'alignement entre chaînes de caractères (Lops *et al.*, 2011). Pour y remédier, les ontologies ont été utilisées dans certains systèmes de recommandation pour servir de support à la représentation des profils utilisateurs (resp. des profils des items). Deux types d'approches sont proposés : l'un consiste à considérer les items comme des instances (Rodríguez-García *et al.*, 2015; Carrer-Neto *et al.*, 2012), l'autre associe un ensemble de concepts/instances de la base de connaissance aux éléments dont on veut définir le profil (items et/ou utilisateurs) (Moreno *et al.*, 2013; Sieg *et al.*, 2010; Middleton *et al.*, 2001).

3.1 Représenter des items à l'aide des instances de concepts

Comme nous l'avons présenté plus haut, la *A-Box* comporte l'ensemble des instances des concepts définis dans une ontologie du domaine ciblé. Il s'agit de l'ensemble des faits ou des objets rattachés à ces concepts lors de la phase de peuplement. Ils sont donc catégorisés selon

la structure de l'ontologie. Pour un SdR qui vise à recommander des items appartenant à un domaine particulier, il est donc possible de rattacher ces items aux éléments de cette *A-Box*. Deux choix sont possibles : i) soit la mise en correspondance est directe et les items sont eux-même assimilés à des instances, ii) soit la mise en correspondance est indirecte et les items sont alors stockés au sein d'un container (*e.g.*, base de données) et possèdent des références liées aux instances.

Le fait de considérer les items au sein d'un SdR comme des instances d'une ontologie du domaine liée à la recommandation possède des avantages significatifs. Une classification des items est systématiquement établie grâce à la structuration des concepts auxquels ils se rattachent. Contrairement à un SdR ne disposant d'aucune information sémantique sur les items, un tel système a l'avantage de disposer de connaissances *a priori* sur les items. Outre certains raisonnements logiques qui peuvent être appliqués grâce aux relations sémantiques du modèle (*i.e.*, inférences), il est également possible d'y appliquer certaines mesures sémantiques et ainsi de considérer que deux items sont proches (resp. éloignés) en fonction des résultats de ces mesures. Ces calculs peuvent être d'une grande finesse si l'ontologie du domaine sur laquelle repose le système est définie de façon détaillée. Ainsi, selon les préférences exprimées par un utilisateur pour un item particulier (ou un ensemble d'items), des items considérés comme étant proches à l'aide de ces mesures peuvent constituer des recommandations judicieuses. De ce fait, on réduit le problème du démarrage à froid pour des nouveaux items dans le contexte d'une approche hybride. En effet, lorsqu'un item n'est encore noté par aucun utilisateur et qu'on ne connaît donc pas le vecteur de notes qui lui est associé, il est tout de même possible de le recommander. Nous détaillons dans la section 4 les mesures de similarité sémantiques des items au sein d'un système de recommandation.

3.2 Modélisation du profil d'utilisateur par des éléments d'une ontologie

Dans une base de connaissance, les instances sont caractérisées par les concepts de l'ontologie auxquels elles sont rattachées et par les attributs liés à ces concepts. Dans le domaine cinématographique, par exemple, *"Bienvenue chez les Ch'tis"* est une instance du concept *Film* possédant des attributs qui le caractérise (*e.g. Genre, Acteurs, Réalisateur, Prix, etc.*). Généralement, lors du peuplement d'une ontologie, on associe les instances aux concepts pertinents qui sont les plus bas dans l'arborescence, *i.e.* les concepts les plus spécifiques. Les liens qui peuvent exister entre une instance et des concepts plus génériques, ancêtres de ceux qui lui sont directement associés ne sont pas explicités car ils peuvent être inférés par le raisonnement.

Un profil d'utilisateur vise à caractériser les préférences de cet utilisateur relativement aux items pour lesquels il s'est déjà exprimé, soit par des préférences explicitées à partir de notes, soit par des interactions avec le système. Si les caractéristiques de ces items sont issues de l'ontologie du domaine de recommandation, le profil utilisateur pourra prendre la forme d'un vecteur dont chaque indice se réfère à un concept (*i.e.* une classe ou un attribut) et dont la valeur représente le taux d'intérêt de l'utilisateur pour ce concept. En d'autres termes, nous pouvons considérer que le profil d'un utilisateur est représenté par les éléments présents dans la *T-Box* associés à différents poids selon les intérêts exprimés. Dans la littérature, cette approche est appelée *profil d'utilisateur ontologique*. Sieg *et al.* (2010) définissent un profil d'utilisateur comme un ensemble de nœuds dont chacun est représenté sous forme d'une paire, $\langle C_j, IS(C_j)\rangle$, où C_j est un concept défini dans l'ontologie et $IS(C_j)$, *i.e. Interest Score*, le taux d'intérêt d'un utilisateur pour C_j. Ils ont employé par la suite ces profils pour estimer la similarité entre des utilisateurs dans une approche hybride. Ce formalisme se base sur le travail de Middleton *et al.* (2004) qui adoptent des triplets, *i.e.* $\langle utilisateur, thème, taux\ d'intérêt\rangle$ pour modéliser l'intérêt des utilisateurs pour des articles scientifiques de différents domaines.

Carrer-Neto *et al.* (2012) proposent un autre moyen de représenter le profil d'un utilisateur, nommé *Recon*, qui repose sur la *A-Box*. Comme précédemment, ce profil est composé d'un vecteur associé à chaque utilisateur, mais dont chaque dimension traduit un degré d'intérêt pour une instance (un item particulier) et non pas un concept. L'avantage d'encapsuler les instances dans le profil utilisateur est que chaque nouvelle évaluation affine la définition du

profil ; Ainsi une part du filtrage est déjà réalisée puisqu'il suffira d'ordonner les composantes du vecteur pour déterminer les items qui ont le plus de poids et en faire la recommandation.

Contrairement aux approches qui adoptent un vecteur (de concepts ou d'instances) pour modéliser le profil d'un utilisateur, (Blanco-Fernandez *et al.*, 2008) présentent un système *CBF* pour recommander des programmes de télévision numérique. La spécificité de leurs travaux est que le profil d'utilisateur est modélisé par un sous-graphe extrait à partir d'une base de connaissance associée à un domaine, *i.e.* celui de la *TV* dans leur cas. Un tel profil d'utilisateur contient des instances (des programmes) pertinentes, leurs principaux attributs et les genres sous lesquels ces programmes sont classés dans l'ontologie. L'intégration des différents types de nœuds dans le profil permettra un filtrage plus fin puisque l'on dispose de plus d'information.

3.3 Inférence du taux d'intérêt et mise à jour du profil utilisateur

L'initialisation des différentes valeurs du vecteur correspondant au degré d'intérêt des utilisateurs est importante puisqu'elle permet de proposer des recommandations dès la première connexion au système. Pour ce faire, en amont de la recommandation, un formulaire ou un questionnaire est souvent soumis à l'utilisateur lors de son inscription (*c.f.* Annexe B). Au lieu de questionner des utilisateurs sur leurs préférences par rapport à chaque concept de l'ontologie pour engendrer un profil initial complet, Moreno *et al.* (2013) proposent dans leur système de recommandation des activités touristiques, de ne questionner l'utilisateur que sur quelques concepts généraux mais suffisamment significatifs pour représenter les principaux centres d'intérêt des touristes (*e.g. Plage, Shopping, Culture, Gastronomie*, *etc.*). L'avantage de demander directement à l'utilisateur d'exprimer ses goûts pour construire son profil est que cela permet d'obtenir des recommandations précises, *e.g.*, s'il aime la plage, on lui recommandera des activités près de la mer. Cependant, cela demande un effort à l'utilisateur, et cette activité en plus d'être chronophage peut également être perçue comme trop intrusive. Paradoxalement, dans la plupart des cas, les utilisateurs cherchent à recevoir des recommandations précises et pertinentes mais sans trop se dévoiler au système, sans être interrogés de façon précise. Dans ce cas, une solution consiste à attribuer initialement le même poids à chaque concept du profil (Sieg *et al.*, 2010). Ce n'est qu'au fur et à mesure de l'utilisation du SdR que le profil de chaque utilisateur sera affiné et que les recommandation gagneront en pertinence.

Pour améliorer la qualité de la recommandation, le profil d'un utilisateur doit être mis à jour à chaque retour fait au système. Ces preuves d'intérêt fournies par l'utilisateur peuvent exprimer ses préférences d'une façon explicite (*e.g.* une note) ou implicite, par exemple un clic, une sauvegarde, un achat, *etc.* Différents types de retour possèdent des significations différentes. Un retour explicite est souvent considéré comme plus fiable qu'un retour implicite, car ce dernier est soumis à l'interprétation automatique de la machine et est donc moins fiable. Afin d'ajuster le profil, on ajoutera ou enlèvera des points à certains attributs du profil pour qu'il reflète au mieux l'ensemble des interactions de l'utilisateur avec le système. Parmi les retours d'information explicites, nous pouvons distinguer l'importance relative des actions en fonction des concepts sur lesquels portent les retours. Par exemple dans les travaux de Carrer-Neto *et al.* (2012), un utilisateur a la possibilité de noter non seulement des films, *i.e.* des instances de la classe de référence *Film*, mais aussi d'autres types d'items (*e.g.*, des acteurs, des réalisateurs, *etc.*) qui peuvent être associés à ces films. Les valeurs du vecteur *Recon* sont ajustées différemment selon le type d'instances sur lequel un utilisateur a fourni son retour. Ainsi, on ajuste le taux d'intérêt pour un film avec une valeur plus légère s'il s'agit d'un retour sur ses acteurs que s'il s'agit d'un retour sur le film lui même.

Dans une ontologie, comme mentionné dans la section 2.2, on dispose des relations hiérarchiques et sémantiques entre les concepts. Elles permettent de relier les concepts et de leur conférer du sens. De ce fait, une propagation des préférences au travers des relations pourra être prise en compte lors de l'ajustement du profil, *e.g.* si on aime des romans de suspense alors on pourra éventuellement aimer ceux de mystère vu que la catégorie suspense est une sous-catégorie de mystère. Pour ce faire, l'algorithme de *Spreading Activation* permet de propager des préférences entre les nœuds via leur relations hiérarchiques (Sieg *et al.*, 2007;

Moreno *et al.*, 2013; Middleton *et al.*, 2001).

4 Mesures de similarité sémantique entre items et/ou entre utilisateurs

« *Si vous aimez l'objet A, vous aimerez l'objet B* », voici un exemple typique de ce que l'on peut lire sur un site de commerce électronique. Derrière cette suggestion se cache un SdR qui nous recommande B parce qu'il a jugé que son contenu est *similaire* à celui de l'objet A que nous venons d'acheter ou de consulter, *i.e.* un SdR de type *CBF*. Cette proximité entre items repose sur une mesure de distance entre leurs profils, qui se résume par conséquent, la plupart du temps, à une distance entre leurs vecteurs de caractéristiques (*e.g.* l'angle, la corrélation, *etc.*). Deux problèmes se posent : i) la sélection des propriétés représentatives est une tâche complexe et souvent les caractéristiques sélectionnées n'arrivent pas à couvrir la totalité des descriptions des items ; ii) il existe souvent des propriétés redondantes parmi les dimensions du vecteur qui peuvent donc biaiser les résultats. Par exemple, dans le domaine cinématographique, les propriétés *suspense* et *mystère* traduisent une granularité différente et sont, par conséquent, en partie redondantes.

L'autre moyen d'apprécier la proximité entre items, adopté par des approches de filtrage collaboratif, consiste à mesurer la distance entre des vecteurs de notes présents dans la matrice d'*utilisateurs-items*. La limite de cette mesure vient du fait que la matrice est généralement creuse et qu'on n'a qu'une vue partielle de la similarité entre items si l'on exploite uniquement leurs notes (deux films peuvent avoir les mêmes notes pour des raisons très différentes).

Comme nous l'avons vu, les items d'un système de recommandation peuvent être représentés par des *instances de concepts* définis dans l'ontologie du domaine ciblé par la recommandation. La mesure de similarité entre ces ressources au travers de leur caractérisation sémantique pourra être adoptée. Cette mesure sémantique s'applique soit d'une façon indépendante (Harispe *et al.*, 2013), soit en combinant d'autres mesures comme mentionné précédemment (Gao *et al.*, 2009; Al-Hassan *et al.*, 2011; Carrer-Neto *et al.*, 2012; Benouaret, 2017).

L'atout majeur de l'estimation des proximités entre items grâce à une mesure sémantique entre instances des concepts d'une ontologie est que, d'une part, on lève le problème de la sélection des propriétés représentatives des items dans le cas d'un SdR *Content-Based*. En effet, étant donné que les propriétés de ces items sont contenues et structurées directement dans le modèle elles peuvent intervenir dans le calcul sans filtrage préalable. De plus, les relations qui les connectent fournissent une sémantique riche qui permet d'interpréter la recommandation produite. D'autre part, utiliser l'ontologie de domaine pour calculer ces similarités pallie le problème du *démarrage à froid* dans le cas de SdR *CF*, car cela permet de proposer une mesure de comparaison solide entre items, même si ceux-ci n'ont été notés que par peu ou pas d'utilisateurs.

Par exemple, pour apprécier la similarité entre deux instances de la classe *Film* dans l'approche proposée par (Carrer-Neto *et al.*, 2012), le calcul se base sur la proportion du nombre d'instances partagées par deux films au vu de l'ensemble des propriétés non taxonomiques (propriétés d'objet ou de données – *object properties* et *datatype properties* en OWL) que la classe de référence *Film* comporte. Formellement, la similarité entre deux instances a et b est mesurée par l'équation (1) où P représente l'ensemble des $\#P$ propriétés de la classe de référence, $common(a, b, P[i])$ représente le nombre d'instances partagées par a et b via la propriété $P[i] \in P$, $deg(a, P[i])$, le nombre d'instances associées à l'individu via la propriété $P[i]$ et $Weight(P[i])$, quant à lui, est une valeur entre 0 et 1 qui indique le poids associé à la propriété, *e.g.* le genre d'un film est certainement plus important que l'endroit où le film a été réalisé. L'avantage de cette approche est sa simplicité de calcul. Son inconvénient est qu'il ne considère que le nombre d'instances partagées par deux items mais pas les contenus (types, valeurs, *etc.*) de ces instances ; or, deux genres de film pourraient être proches même s'il s'agit de deux instances différentes. (Rodríguez-García *et al.*, 2015) étendent cette mesure de similarité par la prise en compte des valeurs des propriétés du type *datatype properties*,

e.g. s'il s'agit de chaînes de caractères alors la distance de *Levenshtein* s'appliquera.

$$S(a,b) = \sum_{i=1}^{\#P} \left(\frac{common(a,b,P[i])}{max(deg(a,P[i]), deg(b,P[i]))}\right) \cdot Weight(P[i]) \quad (1)$$

(Harispe *et al.*, 2013) proposent une mesure de similarité sémantique dédiée à l'estimation des proximités des instances présentes dans une base de connaissance RDF. La notion de projection des instances est introduite dans l'approche pour caractériser leurs propriétés. Une projection représente un ensemble de chemins dont les points de départ sont des instances. Selon l'ensemble d'arrivée de leur dernière propriété, différents types de projections sont ainsi distingués : *donnée*, s'il s'agit d'un ensemble de nœuds du type littéral (*e.g.* valeur numérique, chaîne de caractères, *etc.*) ; *instance*, un ensemble d'instances ; *conceptuel*, un ensemble de concepts (classes en OWL) ; *complexe*, si plusieurs chemins sont nécessaires pour caractériser une propriété. Ainsi, pour mesurer la proximité entre deux instances, on associe à chaque type de projection, une mesure permettant de comparer une paire de projections d'instances. Par exemple, pour une paire de projections de type *conceptuel*, l'approche introduite par (Pesquita *et al.*, 2009) pourra être adoptée.

Dans des approches hybrides, différents types de mesures de similarité peuvent être activés ou désactivés selon le contexte de la recommandation. (Benouaret, 2017) propose un système de recommandation qui vise à suggérer à ses utilisateurs des œuvres à visiter en s'adaptant à leurs préférences. Trois composantes, *i.e.* i) démographique, ii) sémantique et iii) collaborative s'activent consécutivement en fonction du contexte de la recommandation. Plus précisément, étant donné un utilisateur, ses informations démographiques (i), *e.g.* l'âge, le sexe, le pays, sont utilisées dans le cas où il s'agit de sa première authentification pour trouver l'ensemble des utilisateurs qui possèdent un profil démographique similaire afin de procéder à la recommandation. Ensuite, lorsqu'il a exprimé ses préférences par rapport aux items recommandés par la première approche, l'approche sémantique (ii) s'active afin de lui proposer des items qui sont sémantiquement proches de ce qu'il a aimé. Enfin, après avoir collecté suffisamment de notes de la part de l'utilisateur, le filtrage collaboratif (iii), basé sur la similarité des notes, s'active afin de prendre en compte ce qu'ont aimé ses voisins. Notons que pour la composante sémantique (ii) différents calculs s'appliquent pour estimer la proximité entre deux œuvres en fonction des valeurs que prennent leurs propriétés définies dans la base de connaissance. On notera ici l'importance de la qualité des ontologies utilisées. En effet, leur niveau de détail et leur degré de couverture du domaine impactent fortement l'estimation de mesures sémantiques.

Les sections 3 et 4 ont montré comment les ontologies et les bases de connaissance peuvent être employés au sein des systèmes de recommandation. Une vision synthétique de la littérature à ce sujet est proposée en annexe A. On y distingue notamment les domaines d'application concernés.

5 Perspectives de recherche et positionnement

Au vu de l'état de l'art présenté, plusieurs limites des approches usuelles de la recommandation ont été soulevées. Certaines ont fait l'objet de propositions qui intègrent les ontologies de domaines et plus largement les bases de connaissance, pour pallier ces limites. Il nous semble cependant pertinent d'envisager d'aller encore plus loin dans l'exploitation des modèles de connaissance à des fins de recommandation. Cette section propose différentes pistes que nous souhaitons explorer dans le cadre de nos travaux. Il y sera notamment question i) d'intégration des données liées, ii) de diversification des recommandations, iii) de l'explicabilité des recommandations et leur interprétation, et enfin iv) de la recommandation pour un groupe d'utilisateurs. Certaines ont déjà été abordées dans la littérature, d'autres ouvrent la voie à de nouveaux champs exploratoires.

Les données liées, souvent abrégées par LOD pour *Linked Open Data*, représentent un gigantesque graphe de données et de connaissances, présent sur Internet, structuré par de multiples connexions entre ces ressources. Heitmann & Hayes (2010) proposent de tirer parti

des technologies liées à ce paradigme pour enrichir l'approche de filtrage collaboratif lors des démarrages à froid. (Beldjoudi *et al.*, 2016) propose une approche qui explore le LOD afin de réduire le problème de la *sérendipité*. Des travaux récents et similaires comme (Di Noia *et al.*, 2012; Tomeo *et al.*, 2016; Musto *et al.*, 2018) s'inspirent tous des technologies du Web sémantique et des données liées en les combinant avec des technologies innovantes liées à l'apprentissage automatique et l'apprentissage profond.

La diversité des recommandations est un critère crucial à considérer lors de l'évaluation d'un SdR (Gunawardana & Shani, 2015). Elle vient pallier le problème de sur-spécialisation, souvent mentionné dans la littérature (Benouaret, 2017). Ainsi lorsque le SdR recommande des items en relation avec un profil utilisateur, celui-ci se voit recommander des ressources similaires à celles qu'il a déjà aimé, au risque de créer une lassitude, voire d'être totalement inadapté : s'il a acheté un meuble de salle de bain, quel est l'intérêt de lui en proposer d'autres ? Ainsi la diversification des recommandations consiste à proposer à l'utilisateur une liste de recommandations dans laquelle chaque élément diffère au maximum des autres. Par exemple, au lieu de proposer une liste de films de la même catégorie, ou du même réalisateur (sous prétexte que l'utilisateur aime telle catégorie ou tel réalisateur), le but est de rechercher des films en lien avec ses goûts mais qui possèdent le plus de diversité possible. Un des défis dans le domaine consiste à augmenter la diversification des recommandations sans perdre en précision (*i.e.* compromis entre diversité et précision). En effet s'il faut que les items recommandés soient différents pour éviter les redondances, il faut impérativement qu'ils en demeurent pertinents. Cette problématique a fait l'objet de nombreux travaux dans le domaine de la recherche d'information (Gollapudi & Sharma, 2009; Clarke *et al.*, 2011). Des solutions ont été proposées qui mettent en oeuvre les ontologies dans le domaine de la RI (Besbes & Baazaoui-Zghal, 2018). Nous pensons que leur adaptation au contexte des SdR pourrait amener des améliorations significatives. Dans ce cas, les ontologies et les données liées peuvent constituer un support de qualité car nous avons, d'une part, la possibilité de modéliser sémantiquement les goûts d'un utilisateur pour évider la perte de précision et, d'autre part, le raisonnement et la richesse de ressources présentes dans les données liées permettent d'augmenter la diversité.

La façon de présenter les recommandations et de les expliquer doit rendre le processus transparent pour l'utilisateur. En effet il doit comprendre pourquoi des items lui ont été proposés de façon à mieux adapter son interaction avec le système et lui accorder une plus grande confiance (Tintarev & Masthoff, 2015). La recherche dans ce contexte est relativement récente. Des modèles sémantiques très encourageants ont été proposés. Par exemple, une représentation (ou, visualisation) des recommandations sous la forme d'un graphe hiérarchique adossé à une ontologie peut aider l'utilisateur dans sa compréhension des liens sémantiques entre les items proposés et ceux qui ne l'ont pas été (Thanapalasingam *et al.*, 2018) ; le raisonnement peut également être utilisé comme mécanisme d'explication des recommandations.

La dernière piste de recherche que nous proposons concerne la recommandation pour un groupe d'utilisateurs plutôt que pour un utilisateur donné. Certaines propositions ont été faites dans le domaine des SdR (O'Connor *et al.*, 2001; Masthoff, 2010) mais sans prise en compte d'une connaissance *a priori* du domaine ou de la sémantique liée à cette recommandation. Or il nous semble, après nos travaux concernant le clustering sémantique et l'annotation de groupes d'éléments (Fiorini, 2015), que les approches sémantiques offrent un cadre tout à fait propice au développement de méthodes innovantes et performantes dans ce contexte.

6 Conclusion

Cet article propose un état de l'art et un positionnement concernant l'apport des ontologies, et plus largement des bases de connaissance, au domaine des systèmes de recommandation. Une analyse fouillée est présentée. Cet état de l'art illustre la manière dont les ontologies peuvent être intégrées à un système de recommandation pour en améliorer les performances. Deux principaux usages ont été explorés jusque là : l'un vise à représenter les profils des utilisateurs du système, soit par des concepts définis dans l'ontologie, soit par des instances de la classe qui se réfère à la recommandation, les valeurs associées à chaque élément du profil

utilisateur seront ajustées en fonction des actions de l'utilisateur et de ses retours d'information ; l'autre usage des ontologies vise principalement à mesurer la proximité sémantique des items ou des utilisateurs, représentés par des instances des classes d'une ontologie en prenant en compte les propriétés qui les caractérisent et leurs relations sémantiques. Une figure représentant l'intégralité du processus de recommandation et les différents apports que peuvent constituer les ontologies au sein de ce processus est proposée en annexe B.

Parmi les systèmes de recommandation étudiés dans cet article, les ontologies semblent surtout pertinentes dans le cadre d'approches *CBF* ou *Hybrides*, c.f. tableau de synthèse présenté dans l'annexe A. Ceci est dû au fait que dans une approche de filtrage collaboratif, la recommandation ne se joue généralement qu'avec la matrice de notes où aucune information relative aux items n'est prise en compte. Les pistes de recherches proposées ouvrent la voie à de nouvelles avancées du domaine.

Références

AGGARWAL C. C. (2016). *An Introduction to Recommender Systems*, In *Recommender Systems : The Textbook*, p. 1–28. Springer International Publishing : Cham.

AL-HASSAN M., LU H. & LU J. (2011). Personalized e-government services : Tourism recommender system framework. In *Lecture Notes in Business Information Processing*, volume 75 LNBIP, p. 173–187 : Springer, Berlin, Heidelberg.

ALEC C. (2016). *Ontology enrichment and population from texts and data from LOD : Application to automatic annotation of documents*. Thèse de doctorat en informatique, Université Paris-Saclay.

ALI F., ISLAM S. R., KWAK D., KHAN P., ULLAH N., JO YOO S. & KWAK K. (2018). Type-2 fuzzy ontology–aided recommendation systems for iot–based healthcare. *Computer Communications*, **119**, 138 – 155.

BELDJOUDI S., SERIDI H. & BENZINE A. (2016). Améliorer la Recommandation de Ressources dans les Folksonomies par l'Utilisation de Linked Open Data. In *IC2016 : Ingénierie des Connaissances*, Montpellier, France.

BENNETT J. & LANNING S. (2007). The netflix prize. In *Proceedings of the KDD Cup Workshop 2007*, p. 3–6, New York : ACM.

BENOUARET I. (2017). *A contextual and composite recommender system for the personalization of cultural sites visit*. Thèse de doctorat en informatique, Université de Technologie de Compiègne.

BESBES G. & BAAZAOUI-ZGHAL H. (2018). Fuzzy ontologies for search results diversification : Application to medical data. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, SAC '18, p. 1968–1975, New York, NY, USA : ACM.

BLANCO-FERNANDEZ Y., PAZOS-ARIAS J., GIL-SOLLA A., RAMOS-CABRER M. & LOPEZ-NORES M. (2008). Providing entertainment by content-based filtering and semantic reasoning in intelligent recommender systems. *IEEE Transactions on Consumer Electronics*, **54**(2), 727–735.

BROEKSTRA J., KLEIN M., DECKER S., FENSEL D., VAN HARMELEN F. & HORROCKS I. (2002). Enabling knowledge representation on the web by extending rdf schema. *Computer Networks*, **39**(5), 609 – 634.

BURKE R. (2000). Knowledge-based recommender systems. *Encyclopedia of library and information systems*, **69**.

BURKE R. (2002). Hybrid recommender systems : Survey and experiments. *User Modeling and User-Adapted Interaction*, **12**(4), 331–370.

CANTADOR I., BELLOGÍN A. & CASTELLS P. (2008). News@hand : A semantic web approach to recommending news. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5149 LNCS, p. 279–283 : Springer.

CARRER-NETO W., HERNÁNDEZ-ALCARAZ M. L., VALENCIA-GARCÍA R. & GARCÍA-SÁNCHEZ F. (2012). Social knowledge-based recommender system. Application to the movies domain. *Expert Systems with Applications*, **39**(12), 10990–11000.

CHEN R.-C., HUANG Y.-H., BAU C.-T. & CHEN S.-M. (2012). A recommendation system based on domain ontology and swrl for anti-diabetic drugs selection. *Expert Systems with Applications*, **39**(4), 3995 – 4006.

CLARKE C. L., CRASWELL N., SOBOROFF I. & ASHKAN A. (2011). A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, p. 75–84, New York, NY, USA : ACM.

IC 2019

DI NOIA T., MIRIZZI R., OSTUNI V. C., ROMITO D. & ZANKER M. (2012). Linked open data to support content-based recommender systems. In *Proceedings of the 8th International Conference on Semantic Systems*, p. 1–8 : ACM.

DRAGONI M., DA COSTA PEREIRA C. & TETTAMANZI A. G. (2012). A conceptual representation of documents and queries for information retrieval systems by using light ontologies. *Expert Systems with Applications*, **39**(12), 10376 – 10388.

EL-DOSUKY M. A., RASHAD M. Z., HAMZA T. T. & EL-BASSIOUNY A. H. (2012). Food recommendation using ontology and heuristics. In *International conference on advanced machine learning technologies and applications*, p. 423–429 : Springer.

FIORINI N. (2015). *Semantic similarities at the core of generic indexing and clustering approaches*. Thèse de doctorat en informatique, Ecole doctorale I2S, Université de Montpellier.

GAO F., LI Y., HAN L. & MA J. (2009). InfoSlim : an ontology-content based personalized mobile news recommendation system. In *Wireless Communications, Networking and Mobile Computing, 2009. WiCom'09. 5th International Conference on*, p. 1–4 : IEEE.

GOLLAPUDI S. & SHARMA A. (2009). An axiomatic approach for result diversification. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, p. 381–390, New York, NY, USA : ACM.

GRUBER T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, **5**(2), 199 – 220.

GUNAWARDANA A. & SHANI G. (2015). *Evaluating Recommender Systems*, In F. RICCI, L. ROKACH & B. SHAPIRA, Eds., *Recommender Systems Handbook*, p. 265–308. Springer US : Boston, MA.

HARISPE S., RANWEZ S., JANAQI S. & MONTMAIN J. (2013). Mesures sémantiques basées sur la notion de projection RDF pour les systèmes de recommandation. In *24e journées d'ingénierie des connaissances - IC2013*.

HEITMANN B. & HAYES C. (2010). Using Linked Data to Build Open, Collaborative Recommender Systems. In *AAAI spring symposium : linked data meets artificial intelligence*, volume 2010.

IAQUINTA L., D. GEMMIS M., LOPS P., SEMERARO G., FILANNINO M. & MOLINO P. (2008). Introducing serendipity in a content-based recommender system. In *2008 Eighth International Conference on Hybrid Intelligent Systems*, p. 168–173.

LEMDANI R. (2016). *A hybrid and adaptive framework for recommender systems*. Theses, Université Paris-Saclay.

LOPS P., DE GEMMIS M. & SEMERARO G. (2011). *Content-based Recommender Systems : State of the Art and Trends*, In F. RICCI, L. ROKACH, B. SHAPIRA & P. B. KANTOR, Eds., *Recommender Systems Handbook*, p. 73–105. Springer US : Boston, MA.

LU J., WU D., MAO M., WANG W. & ZHANG G. (2015). Recommender system application developments : A survey. *Decision Support Systems*, **74**, 12 – 32.

MASTHOFF J. (2010). *Group recommender systems : combining individual models*, In F. RICCI, L. ROKACH, B. SHAPIRA & P. KANTOR, Eds., *Recommender systems handbook*, p. 677–702. Springer Science+Business Media.

MIDDLETON S. E., ALANI H., SHADBOLT N. R. & DE ROURE D. C. (2002). Exploiting Synergy Between Ontologies and Recommender Systems. In *SemWeb'02 Proceedings of the 3rd International Conference on Semantic Web*, p. 41–50.

MIDDLETON S. E., DE ROURE D. C. & SHADBOLT N. R. (2001). Capturing knowledge of user preferences : ontologies in recommender systems. In *Proceedings of the 1st international conference on Knowledge capture*, p. 100–107 : ACM.

MIDDLETON S. E., SHADBOLT N. R. & DE ROURE D. C. (2004). Ontological user profiling in recommender systems. *ACM Trans. Inf. Syst.*, **22**(1), 54–88.

MORENO A., VALLS A., ISERN D., MARIN L. & BORRÀS J. (2013). SigTur/E-Destination : Ontology-based personalized recommendation of Tourism and Leisure Activities. *Engineering Applications of Artificial Intelligence*, **26**(1), 633–651.

MUSTO C., FRANZA T., SEMERARO G., DE GEMMIS M. & LOPS P. (2018). Deep Content-based Recommender Systems Exploiting Recurrent Neural Networks and Linked Open Data. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, p. 239–244 : ACM.

OARD D. & KIM J. (1998). Implicit feedback for recommender systems. In *Proceedings of the AAAI Workshop on Recommender Systems*, p. 81–83.

OBEID C., LAHOUD I., EL KHOURY H. & CHAMPIN P.-A. (2018). Ontology-based recommender system in higher education. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, p. 1031–1034, Republic and Canton of Geneva, Switzerland : International World Wide Web

Conferences Steering Committee.

O'CONNOR M., COSLEY D., KONSTAN J. A. & RIEDL J. (2001). Polylens : A recommender system for groups of users. In *Proceedings of the Seventh Conference on European Conference on Computer Supported Cooperative Work*, ECSCW'01, p. 199–218, Norwell, MA, USA : Kluwer Academic Publishers.

PESQUITA C., FARIA D., FALCÃO A. O., LORD P. & COUTO F. M. (2009). Semantic similarity in biomedical ontologies. *PLOS Computational Biology*, **5**(7), 1–12.

RICCI F., SHAPIRA B. & ROKACH L. (2015). Recommender systems : Introduction and challenges. In *Recommender Systems Handbook, Second Edition*, p. 1–34. Boston, MA : Springer US.

RODRÍGUEZ-GARCÍA M. Á., COLOMBO-MENDOZA L. O., VALENCIA-GARCÍA R., LOPEZ-LORCA A. A. & BEYDOUN G. (2015). Ontology-based music recommender system. In S. OMATU, Q. M. MALLUHI, S. R. GONZALEZ, G. BOCEWICZ, E. BUCCIARELLI, G. GIULIONI & F. IQBA, Eds., *Distributed Computing and Artificial Intelligence, 12th International Conference*, p. 39–46, Cham : Springer International Publishing.

SARWAR B., KARYPIS G., KONSTAN J. & RIEDL J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, p. 285–295, New York, NY, USA : ACM.

SCHAFER J. B., FRANKOWSKI D., HERLOCKER J. & SEN S. (2007). *Collaborative Filtering Recommender Systems*, In P. BRUSILOVSKY, A. KOBSA & W. NEJDL, Eds., *The Adaptive Web : Methods and Strategies of Web Personalization*, p. 291–324. Springer Berlin Heidelberg : Berlin, Heidelberg.

SCHAFER J. B., KONSTAN J. & RIEDL J. (1999). Recommender systems in e-commerce. In *Proceedings of the 1st ACM Conference on Electronic Commerce*, EC '99, p. 158–166, New York, NY, USA : ACM.

SCHANK R. C., KOLODNER J. L. & DEJONG G. (1981). Conceptual information retrieval. In *Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval*, SIGIR '80, p. 94–116, Kent, UK, UK : Butterworth & Co.

SIEG A., MOBASHER B. & BURKE R. (2010). Improving the effectiveness of collaborative recommendation with ontology-based user profiles. In *proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, p. 39–46 : ACM.

SIEG A., MOBASHER B. & BURKE R. D. (2007). Learning ontology-based user profiles : A semantic approach to personalized web search. *IEEE Intelligent Informatics Bulletin*, **8**, 7–18.

STUDER R., BENJAMINS V. & FENSEL D. (1998). Knowledge engineering : Principles and methods. *Data & Knowledge Engineering*, **25**(1), 161 – 197.

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.

TARUS J. K., NIU Z. & MUSTAFA G. (2018). Knowledge-based recommendation : a review of ontology-based recommender systems for e-learning. *Artificial Intelligence Review*, **50**(1), 21–48.

TARUS J. K., NIU Z. & YOUSIF A. (2017). A hybrid knowledge-based recommender system for e-learning based on ontology and sequential pattern mining. *Future Generation Computer Systems*, **72**, 37–48.

THANAPALASINGAM T., OSBORNE F., BIRUKOU A. & MOTTA E. (2018). Ontology-based recommendation of editorial products. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11137 LNCS, p. 341–358 : Springer, Cham.

TINTAREV N. & MASTHOFF J. (2015). *Explaining Recommendations : Design and Evaluation*, In F. RICCI, L. ROKACH & B. SHAPIRA, Eds., *Recommender Systems Handbook*, p. 353–382. Springer US : Boston, MA.

TOMEO P., FERNÁNDEZ-TOBÍAS I., DI NOIA T. & CANTADOR I. (2016). Exploiting linked open data in cold-start recommendations with positive-only feedback. In *Proceedings of the 4th Spanish Conference on Information Retrieval*, p.Ĩ1 : ACM.

IC 2019

A Catégorisation des différents travaux cités en fonction de leurs usages des ontologies et des domaines d'applications auxquels ils de réfèrent.

Référence	Domaine d'application	Approche du SdR	Apports des ontologies
(Middleton *et al.*, 2001)	articles scientifiques	*CBF*	PU
(Middleton *et al.*, 2002)	articles scientifiques	*Hybride*	PU, MS
(Blanco-Fernandez *et al.*, 2008)	programmes de télévision numérique	*CBF*	PU
(Cantador *et al.*, 2008)	information (news)	*CBF*	PU
(Gao *et al.*, 2009)	information (news)	*CBF*	PU, MS
(Sieg *et al.*, 2010)	livre	*Hybride*	PU
(Al-Hassan *et al.*, 2011)	e-tourisme	*CBF*	MS
(Carrer-Neto *et al.*, 2012)	film	*Hybride*	PU, MS
(El-Dosuky *et al.*, 2012)	nourriture	*CBF*	PU
(Harispe *et al.*, 2013)	musique	*CBF*	MS
(Moreno *et al.*, 2013)	tourisme	*Hybride*	PU
(Tarus *et al.*, 2017)	e-learning	*CF*	MS
(Benouaret, 2017)	tourisme	*Hybride*	PU, MS
(Thanapalasingam *et al.*, 2018)	produit éditorial	*CBF*	PU

Légende :
PU : profil d'utilisateur ;
MS : mesure sémantique ;
CBF : filtrage basé sur le contenu (content based filtering) ;
CF : filtrage collaboratif (collaborative filtering)

B Apport des ontologies dans différentes phases d'un système de recommandation

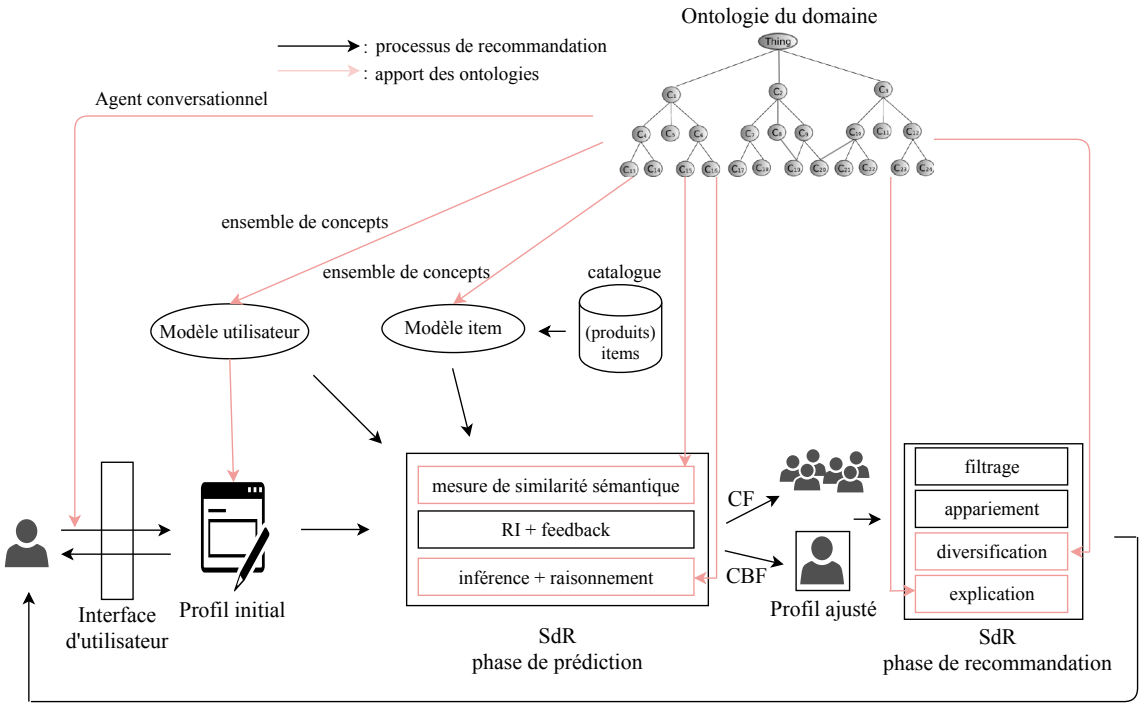


FIGURE *1 – Les parties colorées intègrent une dimension sémantique*

Une nouvelle méthode ensembliste pour la reconnaissance et la désambiguïsation d'entités nommées en utilisant des réseaux de neurones

Lorenzo Canale[1,2], Pasquale Lisena[1], and Raphaël Troncy[1]

[1] EURECOM, Sophia Antipolis, France
{canale|lisena|troncy}@eurecom.fr

[2] Politecnico di Torino, Turin, Italie

Mots-clés : Annotation sémantique, extraction d'entités nommées, désambiguïsation, apprentissage profond.

Résumé de [Canale *et al.* (2018)], publié à ISWC 2018.

Une tâche cruciale en extraction de connaissances à partir de textes se décompose souvent en deux tâches complémentaires : la reconnaissance d'entité nommée (NER) et la désambiguïsation d'entité nommée (NED). L'objectif consiste à attribuer à des parties du texte (mention) respectivement un type appartenant à une taxonomie prédéfinie et un identifiant unique, souvent représenté sous la forme d'URI, qui fait référence de manière univoque à une entité définie dans une base de connaissances donnée. La combinaison de ces deux tâches est souvent abrégée avec l'acronyme NERD.

De nombreuses approches, souvent exposées sous forme d'API Web, ont été proposées pour résoudre ces tâches au cours des dernières années. En termes de NER, chaque service fournit généralement sa propre taxonomie de types qui peuvent être reconnus. Même si tous comprendront trois types principaux (PERSON, ORGANIZATION, LOCATION), ils diffèrent largement pour les types plus fins, ce qui complique leur comparaison et leur combinaison. En termes de NED, chaque extracteur peut potentiellement lever l'ambiguïté d'entités par rapport à des bases de connaissances spécifiques (KB), mais en pratique, ils s'appuient principalement sur des bases de connaissances généralistes, comme DBpedia ou Wikidata. Pour cette raison, la comparaison et la fusion des résultats de ces extracteurs nécessitent certaines tâches de post-traitement qui dépendent généralement d'alignements entre ces bases de connaissances.

Dans ce travail, nous décrivons **Ensemble NERD**, un framework qui regroupe de nombreuses réponses d'extracteurs, les normalise et les combine afin de produire des annotations sémantiques. Cette méthode repose sur deux réseaux d'apprentissage profond, ENNTR (Ensemble Neural Network for Type Recognition) et ENND (Ensemble Neural Network for Disambiguation), qui fournissent des modèles pour effectuer d'une part un alignement entre les types et d'autre part entre les entités nommées identifiées dans une base de connaissances.

En entrée, ces réseaux reçoivent une représentation vectorielle de quatre types de caractéristiques différentes :

- Caractéristiques de forme de surface, liées au texte, à partir desquelles nous avons calculé un plongement lexical ;
- Caractéristiques de type, encodage one-hot calculé sur la taxonomie de chaque extracteur source ;
- Caractéristiques des entités, comparaison des attributs des entités extraites (étiquettes, uris, résumés, etc.)
- Caractéristiques de score, qui incluent certains scores renvoyés par les extracteurs, tels que la saillance ou la confiance.

Chaque type de caractéristique passe auparavant à travers une couche dense qui fonctionne de manière autonome par rapport aux autres, pour leur fournir une couche ensablée, qui a pour résultat la probabilité de correction d'une extraction spécifique. Cette stratégie est évaluée par rapport à des jeux de données bien connus, montrant que la production de l'ensemble surpasse les résultats obtenus par des extracteurs pris individuellement.

IC 2019

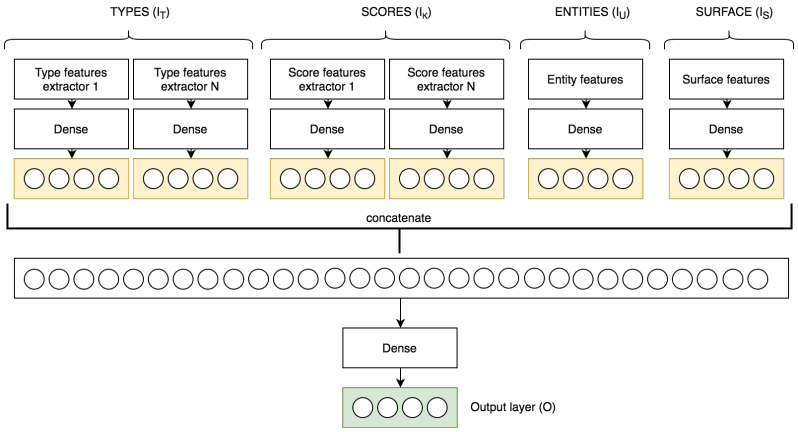


FIGURE *1 – Architecture ENNTR*

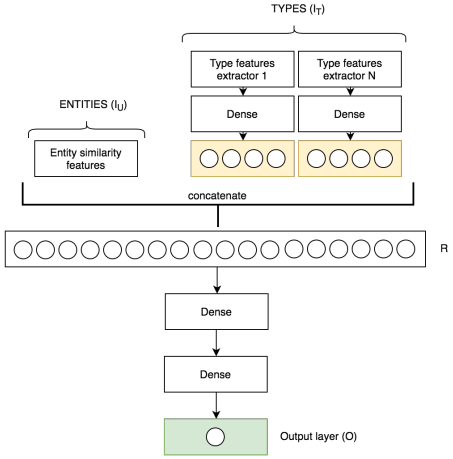


FIGURE *2 – Architecture ENND*

Les modèles produits permettent d'avoir une meilleure précision en termes de mesures micro et macro F1 par rapport aux résultats de chaque extracteur, en utilisant le framework GERBIL. De plus, le réseau ENNTR permet d'éviter l'alignement manuel entre les taxonomies de type de chaque extracteur et la taxonomie cible, en mettant en place un alignement automatique dans le premier niveau du réseau de neurones. Nous avons démontré l'importance de la sélection de caractéristiques (features) pour le succès de ces méthodes ensemblistes. En termes de NER, les formes textuelles jouent un rôle essentiel dans l'ensemble. Pour la tâche NED, si en peut bien dire que les entités aient l'impact le plus important, seule une combinaison avec leurs types permet d'améliorer réellement l'efficacité de la méthode ensembliste en ce qui concerne les prévisions produites par un seul extracteur.

Remerciements. Ce travail a été partiellement financé par l'Agence Nationale de la Recherche (ANR) dans le cadre du projet DOREMUS (ANR-14-CE24-0020), et par le programme de recherche européen H2020 dans le cadre du projet MeMAD (Subvention No. 780069).

Références

CANALE L., LISENA P. & TRONCY R. (2018). A Novel Ensemble Method for Named Entity Recognition and Disambiguation Based on Neural Network. In *17th International Semantic Web Conference (ISWC)*, Monterey, USA.

Représentativité des bases de connaissances avec la loi de Benford généralisée (Résumé)

Arnaud Soulet[1], Arnaud Giacometti[1], Béatrice Markhoff[1], and Fabian M. Suchanek[2]

[1] Université de Tours, LIFAT
firstname.lastname@univ-tours.fr
[2] Telecom ParisTech, LTCI
suchanek@telecom-paristech.fr

Résumé : L'un des succès incontestés du Web sémantique est la construction d'énormes bases de connaissances. Plusieurs travaux récents utilisent ces bases de connaissances pour découvrir de nouvelles connaissances en calculant des statistiques ou en déduisant des règles à partir des données. Par exemple, selon DBpedia, 99% des villes du Yémen ont une population de plus de 1 000 habitants. Nous pourrions donc en conclure que les villes yéménites ont généralement plus de 1 000 habitants. Mais est-ce vrai dans le monde réel ? Naturellement, la fiabilité de cette affirmation dépend de la qualité de la base de connaissances, à savoir de l'exactitude des faits et de leur complétude. Il est bien connu que les bases de connaissances sont très incomplètes. Ce n'est généralement pas un problème en statistique et en apprentissage automatique, où il est rare d'avoir une description complète de l'univers à étudier. La plupart des approches fonctionnent sur un échantillon de données. Dans de tels cas, il est crucial que cet échantillon soit *représentatif* de tout l'univers (ou du moins que le biais de cet échantillon soit connu). Par exemple, le fait que la base de connaissances ne contienne que la moitié des villes du Yémen ne pose pas de problème si leur répartition entre les différentes tailles correspond à peu près à la répartition du monde réel.

Malheureusement, il n'est pas évident a priori de savoir si une base de connaissances est représentative du monde réel. Par exemple, plusieurs bases de connaissances volumineuses, telles que DBpedia ou YAGO, extraient leurs données depuis Wikipedia. Wikipedia, à son tour, est un ensemble de données issues d'une production participative (*crowdsourcing*). Dans le crowdsourcing, les contributeurs ont tendance à indiquer les informations qui les intéressent le plus. En conséquence, Wikipedia présente des biais culturels. De même, Wikidata est le résultat du crowdsourcing et peut présenter des biais similaires. En particulier, il est probable que des pays tels que le Yémen soient moins bien couverts que des pays tels que la France, en raison de la population des contributeurs. Même si les informations contenues dans ces bases de connaissances sont correctes, elles ne sont pas nécessairement représentatives. Si nous savions à quel point une base de connaissances est représentative, nous pourrions alors savoir s'il est raisonnable ou non de l'exploiter pour calculer des statistiques. Une telle indication devrait, par exemple, nous empêcher de tirer des conclusions hâtives sur la répartition de la population dans les villes du Yémen. Mais, comment estimer si une base de connaissances est représentative ou non ?

Soulet *et al.* (2018) proposent d'étudier la représentativité des bases de connaissances à l'aide de la loi de Benford généralisée. Cette loi paramétrée indique la distribution de fréquence attendue par le premier chiffre significatif dans de nombreux jeux de données numériques du monde réel. Nous utilisons cette loi comme référence pour estimer la quantité de données manquante dans la base de connaissances. Plus précisément, nous présentons une méthode pour calculer une borne inférieure pour le nombre de faits manquants pour qu'une relation soit représentative. Cette méthode fonctionne dans un contexte supervisé (où la relation est connue pour satisfaire la loi de Benford généralisée) et dans un contexte non supervisé (où le paramètre de la loi doit être déduit des données). Nous prouvons que, sous certaines hypothèses, les bornes inférieures calculées sont correctes aussi bien pour le contexte supervisé que non supervisé. Nous montrons avec des expériences sur de véritables bases de connaissances que notre méthode est efficace à la fois pour les contextes supervisés et non supervisés. La méthode non supervisée, en particulier, nous a permis d'auditer 63% des faits de DBpedia.

Mots-clés : base de connaissances ; représentativité ; complétude ; loi de Benford généralisée

Références

SOULET A., GIACOMETTI A., MARKHOFF B. & SUCHANEK F. M. (2018). Representativeness of knowledge bases with the generalized benford's law. In *International Semantic Web Conference*, p. 374–390 : Springer.

Vers une exploitation efficace de grandes bases de connaissances par des graphes de contexte

Nada Mimouni, Jean-Claude Moissinac

TÉLÉCOM PARISTECH, LTCI Paris, France
nada.mimouni@telecom-paristech.fr, jean-claude.moissinac@telecom-paristech.fr

Résumé : Un problème lié à l'exploitation de graphe de connaissances, en particulier lors de traitements avec des méthodes d'apprentissage automatique, est le passage à l'échelle. Nous proposons ici une méthode pour réduire significativement la taille des graphes utilisés pour se focaliser sur une partie utile dans un contexte d'usage donné. Nous définissons ainsi la notion de graphe de contexte comme un extrait d'une ou plusieurs bases de connaissances généralistes (tels que DBpedia, Wikidata, Yago) qui contient l'ensemble d'informations pertinentes pour un domaine spécifique tout en préservant les propriétés du graphe d'origine. Nous validons l'approche sur un extrait de DBpedia pour des entités en lien avec le projet Data&Musée et le jeu de référence KORE selon deux aspects : la couverture du graphe de contexte et la préservation de la similarité entre ses entités.
Mots-clés : Base de connaissances, Graphe de contexte, Similarité, DBPedia, Base Joconde

1 Introduction

Les développements du web sémantique et des données liées (le Linked Open Data ou LOD) au cours de la dernière décennie ont permis la publication et le liage de plusieurs données structurées sur le Web. Le *LOD cloud* [1] est passé de 12 à 1378 ensembles de données et de 500 millions à plus de 130 milliards de triplets RDF entre 2007 et 2019. Ces données concernent plusieurs domaines comme la culture, les sciences du vivant, les données de gouvernements ou les données géographiques. Ce développement a démontré l'intérêt de relier un jeu de données d'un domaine applicatif restreint avec un ensemble de données externes afin d'en tirer une meilleure compréhension et exploitation.

Dans le cadre du projet Data&Musée, un projet exploratoire visant à améliorer les systèmes d'information d'institutions culturelles, nous faisons l'hypothèse que la promotion du patrimoine culturel peut bénéficier des techniques récentes de représentation et d'exploration des connaissances. Pour cela, nous avons besoin d'enrichir des ensembles de données relatives au domaine du projet et d'en assurer l'interopérabilité pour parvenir à accroître la visibilité et l'accessibilité par un plus large public. Une des conséquences directes est l'amélioration des conditions financières des institutions culturelles afin d'assurer une meilleure préservation de ce patrimoine.

Une approche reconnue pour assurer ce type d'exploitation de données est l'utilisation des technologies du web sémantique, qui ont montré leur puissance pour le développement des connaissances dans divers domaines comme le tourisme (Soualah Alila *et al.*, 2016; Al-Ghossein *et al.*, 2018), les villes intelligentes (Consoli *et al.*, 2015; Gyrard *et al.*, 2016) ou la valorisation du patrimoine culturel (Lodi *et al.*, 2017). Ces techniques permettent d'assurer une représentation unifiée de données hétérogènes mais apparentées qui facilite le liage et l'enrichissement de ces données.

Les grandes bases de connaissances comme Yago, DBPedia, DBPedia-Fr et Wikidata sont des ressources très utiles car elles fournissent un stock de connaissances encyclopédiques semi-structurées sur les principes du LOD. Mais ces ressources posent plusieurs problèmes d'exploitation : problèmes d'accès, problèmes de performances, limites sur les usages, etc. qui sont principalement liés à leur très grande taille. Nous nous intéressons ici aux problèmes d'échelle et de performance qui peuvent se poser lorsqu'on veut exploiter des liens vers ces grands graphes de connaissances.

1. https://lod-cloud.net/

Dans cet article nous proposons une représentation alternative simplifiée, fidèle et plus accessible d'une base de connaissance, a fortiori étendue, au travers d'un graphe de contexte. L'algorithme d'extraction que nous proposons construit un graphe de contexte pour un domaine donné, défini par un ensemble d'entités représentatives. Le graphe construit se caractérise par la préservation, dans le cadre de ce domaine, des propriétés du graphe d'origine tout en limitant les problèmes de performance et de passage à l'échelle.

Nous évaluons les propriétés du graphe extrait selon deux critères : sa couverture du domaine et son impact sur les résultats d'un ensemble de mesures de similarité entre les entités extraites. En effet, évaluer la similarité entre les ressources est crucial pour plusieurs applications guidées par les données, telles que la découverte de liens, le *clustering* ou le classement.

Nous avons effectué une série de tests pour valider notre méthode sur des données d'institutions culturelles issues du projet Data&Musée. Les résultats montrent que l'utilisation de graphes de contexte rend l'exploitation de grandes bases de connaissances plus maniable et efficace tout en préservant des propriétés du graphe initial.

Dans ce qui suit, la section 2 présente un état de l'art sur l'utilisation des contextes avec des bases de connaissances. La section 3 rappelle les notions de base sur les graphes sémantiques, donne les définitions que nous utilisons dans notre approche et décrit le processus de construction du graphe de contexte. La section 4 fait une revue des mesures de similarité sur des graphes de connaissances et présente notre mesure définie pour la validation d'un graphe de contexte. Les sections 5 et 6 décrivent les expérimentations et les tests de validation effectués respectivement sur les données de Paris Musées et sur le jeu de données de référence KORE. La conclusion et les perspectives sont données dans la section 7.

2 Travaux connexes

La notion de contexte a été utilisée dans plusieurs travaux basés sur le web sémantique pour différentes applications comme le calcul de similarités entre entités ou entre documents, la découverte des liens d'identités pour le liage des données sur le LOD ou la transformation vectorielle des graphes pour application à des méthodes d'apprentissage automatique (Shen *et al.*, 2015; Raad *et al.*, 2017; Beek *et al.*, 2016; Benedetti *et al.*, 2019; Shi *et al.*, 2017; Luo *et al.*, 2015). Ces approches utilisent un extrait des bases de connaissances, appelé contexte, qui le considèrent comme une partie du grand graphe porteuse de sémantique pour une ou plusieurs ressources.

Sémantique du contexte

Dans (Hulpus *et al.*, 2013), les auteurs décrivent un concept d'intérêt (*concept of interest*) C dans DBpedia par un graphe appelé graphe de sens (*sense graph*) ayant comme racine C. Ils proposent une solution au problème d'étiquetage automatique de thèmes (*automatic topic labelling*) utilisant DBpedia. Les thèmes sont extraits par une méthode de *probabilistic topic modelling* (comme LDA). Pour chaque concept C_i associé à un terme d'un thème identifié, ils extraient un *sense graph* G_i en interrogeant tous les nœuds situés à au plus deux sauts (2-hop) de C_i en prenant en compte récursivement tous les liens de type `skos:broader`, `skos:broaderOf`, `rdfs:sub-ClassOf`, `rdfs:type` et `dcterms:subject`. Les graphes G_i sont ensuite fusionnés pour obtenir le graphe de thème (*topic graph*) G. Dans la même direction, les auteurs dans (Raad *et al.*, 2017; Beek *et al.*, 2016) montrent que l'utilisation de contextes permet de mieux décrire les entités pour les lier via des liens d'identité de type `owl:sameAs`. Un lien d'identité est valide dans un contexte, correspondant à un sous-ensemble de propriétés, si deux instances i_1 et i_2 ont les mêmes valeurs de ces propriétés. Ils postulent que deux instances similaires dans un contexte peuvent ne pas l'être dans un autre avec un sous-ensemble différent de propriétés. Ils montrent ainsi l'importance de la prise en compte du contexte pour le calcul de similarité.

La notion d'extrait d'une base de connaissance a été également étudiée dans des domaines plus spécifiques comme l'IoT ou l'environnement pour réduire la complexité de manipulation des données dans ces domaines. (Gyrard *et al.*, 2016) propose le système LOV4IoT pour la

construction d'applications sémantiques de web d'objets utilisant des ontologies du domaine afin de réduire les espaces de recherche et faciliter l'interrogation. Les résultats dans (Wannous *et al.*, 2017) montrent l'impact positif des optimisations, telles que des contraintes de domaine et des raffinements de voisinage, sur la réduction de la complexité du mécanisme d'inférence sur des bases de connaissance de comportements d'animaux. Ces optimisations ont permis de réduire à moitié le temps de calcul et d'améliorer ainsi le passage à l'échelle.

Similarité dans un contexte

La plupart des méthodes qui comparent des ressources, par exemple en terme de similarité, dans le web sémantique se basent sur un ensemble pré-sélectionné de triplets (Colucci *et al.*, 2016). Pour leur méthode de définition et de calcul du LCS (*Least Common Subsumer* : l'ancêtre taxonomique le plus spécifique qui subsume deux ressources) dans des graphes RDF, les auteurs montrent qu'il est important d'expliciter le sous-graphe du web sémantique qui sert de contexte au calcul de LCS pour une ressource r. Le contexte de r, appelé *rooted r-graph*, est constitué d'un ensemble T_r de triplets tel que toutes les ressources dans T_r sont connectées à r par un chemin dans le graphe RDF. Dans (Cheniki *et al.*, 2016), une mesure de similarité entre entités basée sur le LOD est définie sur un contexte (voisins à une profondeur N) extrait à partir de l'ensemble des données disponibles. Pour le calcul de similarité entre documents, les auteurs dans (Benedetti *et al.*, 2019) définissent le contexte sémantique d'analyse extrait d'une base de connaissances comme DBpedia. A partir de ce contexte, ils créent un vecteur sémantique de contexte qui permet de surpasser les méthodes classiques de similarité inter-documents. Dans (Bhatt *et al.*, 2019), les auteurs décrivent un algorithme de détection et de caractérisation de communautés basé sur les graphes de connaissances. Ils abordent le problème de trouver le contexte qui résume le mieux les nœuds des communautés. L'algorithme utilise une mesure de similarité qui intègre les attributs des nœuds décrits dans des graphes de connaissances hiérarchiques (HKG) spécifiques à un domaine. Ces graphes fournissent des informations pertinentes pour un groupe d'objets du monde réel.

Apprentissage sur contexte

L'utilisation des graphes de connaissances avec des méthodes d'apprentissage automatique a été principalement favorisée par le développement de techniques de transformation vectorielle de graphes (*graph embedding*). Cette transformation préserve les propriétés pertinentes du graphe d'origine comme la topologie (proximité entre voisins) ou la sémantique. Dans ce cadre, (Shi *et al.*, 2017) et (Luo *et al.*, 2015) proposent un *embedding* de graphe de connaissances qui crée des vecteurs mieux représentatifs des entités. L'approche tient compte des contextes explicites (liens entrants et sortants et chemins entre paires d'entités) et implicites (motifs de connectivité contextuelle) entre entités non connectées dans ce graphe. Un contexte implicite est construit à partir de l'hypothèse que les entités connectées à un même noeud sont généralement implicitement liées les unes aux autres, même si elles ne sont pas directement liées dans le graphe.

Taille du contexte

La taille du contexte est un paramètre qui a été discuté dans plusieurs travaux. Dans leur travail sur l'étiquetage automatique de thèmes avec DBpedia (Hulpus *et al.*, 2013), les auteurs utilisent une distance de 2 sauts à partir du nœud de départ. Cette distance a été choisie suite à une série de tests sur l'expansion de nœuds qui a montré qu'à partir de 3 sauts, l'expansion produit des graphes très larges et introduit beaucoup de bruit. Pour la définition d'une mesure de similarité entre entités basée sur le LOD (Cheniki *et al.*, 2016), les auteurs se limitent à des chemins de longueur 2 pour récupérer toutes les ressources équivalentes possibles et enrichir l'espace d'instantiation d'une ressource dans le LOD.

Ces travaux mettent l'accent sur l'intérêt d'utiliser des contextes. Cependant, dans ces approches, l'intégralité de la base est considérée pour calculer le contexte à la volée au moment

de l'utilisation des ressources ce qui pose des problèmes d'accès liés à la taille de la base. Dans notre approche, nous proposons de construire *a-priori* un graphe de contexte, unique pour toutes les ressources d'un domaine, qui servira comme point d'accès optimisé pour les différents traitements dans une application donnée.

3 Graphe de contexte guidé par le domaine

3.1 Rappels sur les graphes sémantiques

Notre approche s'appuie sur des bases de connaissances décrites par une ontologie en OWL et des données représentées en RDF. Une base de connaissance correspond à un schéma conceptuel et un ensemble de faits (déclarations).

Définition. Ontologie. Une ontologie $\mathcal{O}$ correspond à la partie conceptuelle de la base (schéma) qui structure les connaissances dans un domaine donné. Elle peut être représentée par un triplet $\mathcal{O} = (C, P_r, A)$ où C est l'ensemble de classes (concepts d'un domaine), P_r est l'ensemble de propriétés de classes et A est l'ensemble d'axiomes, qui précisent des contraintes sur les propriétés d'une classe. Dans la suite, nous parlerons aussi de *T-Box* pour désigner cette partie conceptuelle des connaissances (voir figure 1).

Définition. Faits et graphe de connaissances. Un graphe de connaissances $\mathcal{KG}$ est défini par un ensemble de faits. Un fait est représenté par un triplet de la forme $\langle sujet, predicat, objet \rangle$. $sujet$ désigne un élément sur lequel on veut affirmer une connaissance ; $predicat$ désigne une propriété qu'on veut associer au $sujet$; $objet$ est la valeur que prend la propriété pour ce $sujet$. L'ensemble des faits constitue la *A-Box* d'un graphe (voir figure 1).

Cette définition fait de $\mathcal{KG}$ un graphe étiqueté orienté où $\mathcal{V}$ est l'ensemble de nœuds (sommets) et $\mathcal{E}$ est l'ensemble des liens entre deux nœuds, liens étiquetés par un prédicat (arête). Un fait décrit par $\langle sujet, predicat, objet \rangle \in \mathcal{E}$ est tel que $sujet, objet \in \mathcal{V}$ et $predicat \in \mathcal{P}$, un ensemble de prédicats, par exemple choisi dans un ensemble de propriétés P_r définis dans une ontologie.
$\mathcal{V}$ est l'union de trois ensembles disjoints :

$$\mathcal{V} = \{v \mid v \in \mathcal{U} \cup \mathcal{B} \cup \mathcal{L}\}$$

où $\mathcal{U}$ = ensemble des URIs (identifiants uniques de ressources),
$\mathcal{B}$ = ensemble des *blank nodes*, des sommets qui ont un rôle technique pour regrouper des propriétés sans les associer à une URI,
$\mathcal{L}$ = ensemble des valeurs littérales ; il s'agit de valeurs typées : chaînes de caractères, valeurs numériques, dates, etc.

Un prédicat lie deux URIs ou *blank nodes* ou une URI ou un *blank node* avec un littéral.

$$\mathcal{E} = \{(v_1, p, v_2) \mid v_1 \in \mathcal{U} \cup \mathcal{B}, v_2 \in \mathcal{U} \cup \mathcal{B} \cup \mathcal{L}, p \in \mathcal{P}\}$$

Dans nos expérimentations, nous utilisons la version française de DBpedia comme base de connaissance généraliste, en raison de sa large couverture et de l'abondance et la diversité des liens qu'elle contient et du fait quelle est reliée à de nombreuses autres bases. La figure 1 montre un exemple d'un sous-graphe $\mathcal{KG}$ de DBpedia.

Définition. Chemin. Un chemin $\mathcal{C}$ de longueur N, $\mathcal{C} = (e_i)_{1 \leq i \leq N}$, est une suite finie non vide de liens de $\mathcal{E}$, avec $N \in \mathbb{N}$, tel que deux liens consécutifs sont adjacents. Deux liens l_1 et l_2 sont adjacents lorsqu'ils partagent un nœud n destination pour l_1 et origine pour l_2.

Définition. Prédicat exclu. On définit un ensemble de prédicats qui seront exclus des chemins construits sur un graphe $\mathcal{KG}$; on note cet ensemble par $\overline{\mathcal{P}}$ tel que $\overline{\mathcal{P}} \subset \mathcal{P}$. L'ensemble

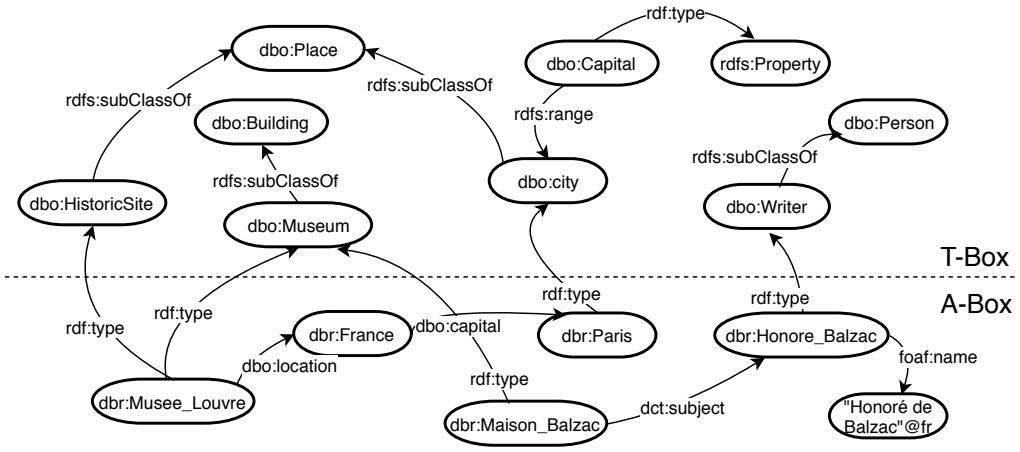


FIGURE *1 – Exemple d'un sous-graphe $\mathcal{KG}$ de DBpedia.*

des liens e étiquetés par des prédicats $p \in \overline{\mathcal{P}}$ est noté par $\overline{\mathcal{E}}$ tel que $\overline{\mathcal{E}} \subset \mathcal{E}$.

Définition. Nœud terminal. Un nœud terminal est un nœud sur lequel on impose l'arrêt de la construction d'un chemin (comme si ce nœud n'avait aucun lien sortant). Un chemin $\mathcal{C}$ construit sur un graphe $\mathcal{KG}$ s'arrête s'il rencontre un nœud terminal ; on note cet ensemble de nœuds par $\overline{\mathcal{V}}$ tel que $\overline{\mathcal{V}} \subset \mathcal{V}$.

3.2 Graphe de contexte

Étant donné un domaine d et un graphe $\mathcal{KG}$, notre objectif est d'extraire un sous-graphe de $\mathcal{KG}$ contenant toutes les informations sur le domaine d qu'on note $\mathcal{CG}(d)$. Notre méthode repose sur les parties *T-Box* et *A-Box* de $\mathcal{KG}$ en considérant des nœuds $v \in (\mathcal{U} \cup \mathcal{L})$ et des liens $e \in (\mathcal{E} \setminus \overline{\mathcal{E}})$.

3.2.1 Définition des paramètres

Pour réduire la taille du graphe de contexte, nous tenons compte d'observations faites sur $\mathcal{KG}$ et de connaissances expertes, quand elles sont disponibles, quant à l'utilité ou l'inutilité de certains nœuds et prédicats.

Plus précisément, la liste $\overline{\mathcal{V}}$ est définie à partir de $\mathcal{V}$ par l'exclusion automatique des nœuds qui appartiennent à la *T-Box* du fait de leur caractère très général (ex. dans DBpedia : `dbo:Building`, `dbo:Place` ou `owl:Thing`) et les nœuds de structuration (ex. mise en forme des pages de DBpedia, `<http://fr.dbpedia.org/resource/Modèle:P.>`).

Parallèlement, la liste $\overline{\mathcal{E}}$ est définie à partir de $\mathcal{E}$ par l'exclusion des liens étiquetés par deux types de prédicats : les prédicats considérés comme peu ou pas informatifs pour le domaine d (liste alimentée par un expert) et les prédicats de structuration de la base $\mathcal{KG}$ (ex. `<http://dbpedia.org/ontology/wikiPageRevisionID>`).

Ces nœuds et prédicats introduisent du bruit sans apporter d'information pertinente pour le domaine considéré. A titre d'exemple, dans nos expérimentations sur DBpedia, nous avons constaté 3486 nœuds construits sur `<http://fr.dbpedia.org/resource/Modele:????>` donnant lieu a 2692515 liens et 101235 liens vers `<http://www.w3.org/2004/02/skos/core#Concept>`. Dans le cadre de notre

application, nous avons construit une liste de nœuds et prédicats exclus qui peut être réutilisée pour des applications dans d'autres domaines. La liste est disponible ici [2].

3.2.2 Processus de construction

L'extraction d'un graphe de contexte $\mathcal{CG}(d)$ de dimension N est un processus récursif qui s'arrête quand la limite N est atteinte. Les étapes du processus sont :

— Identification des germes : une liste de germes $\mathcal{S}(d)$ est définie pour un domaine d'application d. Dans certains domaines cette liste est évidente comme pour le cas des musées, hôtels ou restaurants. Dans le cas général, la pratique commune consiste à se référer à un jeu de données de référence (comme IMDB pour le domaine du cinéma).
— Construction des contextes voisins : un contexte voisin $\mathcal{CV}(a)$ est généré pour toute entité a de $\mathcal{S}(d)$ et la liste des germes $\mathcal{S}(d)$ est mise à jour avec les voisins récoltés.
— Construction du graphe de contexte : le graphe de contexte $\mathcal{CG}(d)$ est construit comme l'agrégation de tous les contextes voisins $\mathcal{CV}$ des germes.

1. Identification des germes. Les germes $\mathcal{S}(d)$ sont des nœuds de $\mathcal{KG}$ qui constituent les entités de départ pour la construction du graphe de contexte $\mathcal{CG}$, $\mathcal{S}(d) = \{v | \forall v \in \mathcal{S}(d), v \in (\mathcal{V} \setminus \overline{\mathcal{V}})\}$. La liste $\mathcal{S}(d)$ est définie pour un domaine d comme l'ensemble des instances de concepts représentatifs du domaine. Par exemple, dans notre cas (projet Data& Musée), les entités de départ correspondent à liste des musées et monuments de *Paris Musées* et du *Centre des Monuments Nationaux*.

Exemple. Sur la figure 1 : $\mathcal{S}(d)$ = { `dbr:Musee_Louvre`, `dbr:Maison_Balzac` }

2. Construction des contextes voisins. Un contexte voisin $\mathcal{CV}$ d'une entité est son voisinage direct (1-hop) dans $\mathcal{KG}$. C'est la structure locale qui interagit avec l'entité et reflète divers aspects de cette entité. Plus précisément, étant donné une entité $a \in \mathcal{S}(d)$, le contexte voisin de a est définit comme suit :

$$\mathcal{CV}(a) = \mathcal{CS}(a) \cup \mathcal{CE}(a)$$

où

$$\mathcal{CS}(a) = \{(a, p, o) \mid \forall (a, p, o) \in \mathcal{E}, \forall p \in (\mathcal{P} \setminus \overline{\mathcal{P}}), o \in (\mathcal{U} \cup \mathcal{L})\}$$

$$\mathcal{CE}(a) = \{(s, p, a) \mid \forall (s, p, a) \in \mathcal{E}, \forall p \in (\mathcal{P} \setminus \overline{\mathcal{P}}), s \in \mathcal{U}\}$$

avec $\mathcal{CS}$ un ensemble de liens sortants de a tandis que $\mathcal{CE}$ est un ensemble de liens entrants de a et s ou o est le nœud voisin de a par un lien étiqueté par p. Notons ici que $a \notin \overline{\mathcal{V}}$, nous ne construisons pas de contextes voisins pour les nœuds terminaux. La liste des germes $\mathcal{S}(d)$ est par la suite mise à jour avec les voisins o et s récoltés tel que $o, s \notin \overline{\mathcal{V}}$.

Exemple. Sur la figure 1 :

$\mathcal{CV}$(`dbr:Musee_Louvre`) = { (`dbr:Musee_Louvre, rdf:type, dbo:HistoricSite`), (`dbr:Musee_Louvre, rdf:type, dbo:Museum`), (`dbr:Musee_Louvre, dbo:location, dbr:France`)}

$\mathcal{CV}$(`dbr:Maison_Balzac`) = { (`dbr:Maison_Balzac, rdf:type, dbo:Museum`), (`dbr:Maison_Balzac, dct:subject, dbr:Honore_Balzac`)}

3. Construction du graphe de contexte. La construction d'un graphe de contexte est un processus récursif, sur la base de l'étape suivante :

2. `https://gitlab.com/snippets/1844328`

$\mathcal{CG}(d)$ pour un domaine d est construit comme l'agrégation de tous les contextes voisin des nœuds germes $a \in \mathcal{S}(d)$ tel que $\mathcal{S}(d) \subset (\mathcal{V} \setminus \overline{\mathcal{V}})$,

$$\mathcal{CG}(d) = \bigcup_{a \in \mathcal{S}(d)} \mathcal{CV}(a)$$

A la fin d'une étape, la liste des germes $\mathcal{S}(d)$ est mise à jour avec les voisins v récoltés tel que $v \notin \overline{\mathcal{V}}$.

Le processus est répété N fois pour un graphe de contexte de profondeur N. Comme démontré par les travaux dans la section 2, $N = 2$ est la valeur la plus intéressante pour un contexte. En effet, la taille du graphe augmente exponentiellement en fonction de la profondeur (pour des entités qui possèdent en moyenne x voisins, à 1-hop la taille est x, à 2-hop la taille est x^2, à 3-hop la taille est x^3, etc.), aller au delà de 2 augmente significativement l'espace et introduit beaucoup de bruit. Dans le cadre de notre expérimentation, nous arriverions au niveau 3 à un graphe du même ordre de grandeur que tout DBPedia.

A la fin du processus, $\mathcal{CG}(d)$ est complété avec la partie *T-Box* de $\mathcal{KG}$ (ici l'ontologie DBpedia [3]) et pour tout nœud dans $\mathcal{CG}(d)$, on assure qu'un lien de type `is-a` existe avec un concept de la partie *T-Box* (si ce lien existe dans le graphe d'origine $\mathcal{KG}$). Le **cœur du contexte** est le graphe obtenu au niveau $N - 1$. Les entités ajoutées au niveau N sont la **périphérie du contexte**.

3.3 CONTEXT : Algorithme de construction d'un graphe de contexte

L'algorithme `CONTEXT` (algorithme 1) construit un graphe de contexte $contexte$ à partir d'un graphe de connaissances $\mathcal{KG}$ pour un domaine d. Pour un ensemble d'entités représentatives d'un domaine, les germes ($germesATraiter$), `ContexteVoisin`(g) extrait un contexte voisin C_v à partir d'un graphe de connaissances $\mathcal{KG}$ pour chaque germe g. Le contexte final, $contexte$, est enrichi par C_v. Une liste de nouveaux germes, $nouveauxGermes$, est mise à jour avec les nouvelles entités récoltées après filtrage des nœuds terminaux avec la méthode `EntitésFiltrées`. La profondeur d'exploration $niveau$ est incrémenté de 1 à chaque étape jusqu'à la limite $rayon$ souhaitée. A la fin du processus, le contexte résultant $contexte$ est enrichi par les classes de l'ensemble de ces entités extraites de $\mathcal{KG}$ par les méthodes `AjoutClasses` et `Entités`.

4 Validation de graphes de contexte par mesure de similarité

4.1 Hypothèse de pertinence d'un graphe de contexte

L'utilisation d'un graphe de contexte construit à partir d'un grand graphe de connaissances permet, comme évoqué plus haut, de gagner en performance (temps de calcul, espace mémoire, etc.). Ce gain ne doit pas par ailleurs pénaliser son utilisation par les méthodes habituellement basées sur la structure et le contenu du graphe d'origine. En effet, plusieurs algorithmes utilisent les graphes de connaissances comme structure de base ou comme source d'enrichissement sémantique pour effectuer plusieurs tâches dans différents domaines tels que l'analyse de réseaux sociaux (e.g. détection de communautés) (Bhatt *et al.*, 2019), la recommandation (e-commerce, tourisme, musique) (Oramas *et al.*, 2016), etc.

La plupart de ces méthodes se base sur la notion de similarité sémantique (*semantic similarity*) ou de proximité sémantique (*semantic relatedness*) entre entités (instances de classes) pour effectuer des traitements finaux sur les données d'origine. Le calcul de ces mesures a plusieurs applications directes et pertinentes pour le traitement automatique de langue (la désambiguïsation, l'annotation sémantique, la recherche d'information, etc.), la découverte

3. `https://wiki.dbpedia.org/services-resources/ontology`

Algorithme 1 : CONTEXT

Fonction *ContexteConstructeur(*, $germesATraiter$, $rayon$, $filtre$*)*
  Entrée : Un graphe de connaissances KG
  Une profondeur $rayon$ de voisinage à atteindre
  Un ensemble d'entités qui servent de germes $germesATraiter$
  Un ensemble d'entités qui ne doivent pas servir de germes $filtre$
  Sortie : Graphe de contexte $contexte$
  $niveau \leftarrow 0$
  $contexte \leftarrow \emptyset$
  tant que $niveau < rayon$ **faire**
    $nouveauxGermes \leftarrow \emptyset$
    pour chaque $g \in germesATraiter$ **faire**
      $C_v \leftarrow$ `ContexteVoisin(`KG, g`)`
      $contexte \leftarrow contexte \cup C_v$
      $nouveauxGermes \leftarrow nouveauxGermes \cup$ `EntitésFiltrées(`C_v, $filtre$`)`
    fin
    $niveau \leftarrow niveau + 1$
    $germesATraiter \leftarrow nouveauxGermes$
  fin
  $contexte \leftarrow contexte \cup$ `AjoutClasses(`KG, `Entités(`$contexte$`))`
  retourner $contexte$
fin

de liens ou le classement.

Partant des cas d'utilisation cités plus haut, nous considérons qu'une mesure de similarité est une condition nécessaire pour évaluer si l'utilisation d'un graphe de contexte peut suffire pour satisfaire les besoins en calcul des tâches relatives aux méthodes appliquées aux graphes d'origine. Nous parlons alors de pertinence d'un graphe de contexte.

Définition. Graphe de contexte pertinent. Un graphe de contexte est dit pertinent pour un domaine donné s'il préserve des propriétés du graphe d'origine pour ce domaine évalués en terme de similarité entre entités.

Hypothèse. Nous faisons l'hypothèse que notre graphe de contexte est pertinent pour un domaine si les similarités relatives de deux entités par rapport à une troisième dans le graphe d'origine, sont préservées dans le graphe de contexte.

Dans les graphes de connaissances, la sémantique qui décrit les ressources est codée selon différents aspects comme par exemple les voisins ou la hiérarchie de classes. Une grande majorité des mesures de similarité existantes considèrent les aspects de manière isolée ce qui ne permet pas de couvrir l'ensemble des propriétés de ces ressources. Nous définissons dans ce qui suit une mesure de similarité plus générale (section 4.3) qui compose l'aspect structurel (les liens de hiérarchie taxonomique) et sémantique (l'ensemble des prédicats) décrivant une entité. Ces mesures seront utilisées dans les sections 5 et 6 afin d'évaluer la pertinence d'un graphe de contexte.

4.2 Revue des mesures de similarité basées sur des graphes de connaissances

Dans la littérature, nous distinguons trois grandes familles de mesures de similarité qui se basent sur les ontologies, la partie T-Box d'une base de connaissance (pour une revue dé-

taillée voir (Sánchez *et al.*, 2012; Harispe *et al.*, 2014)).

Mesures basées sur les liens (*edge-counting*). Ces mesures utilisent le nombre de liens séparant les nœuds (relation transversale) comme critère de similarité. La mesure la plus directe est celle définie par (Rada *et al.*, 1989) qui calcule le chemin C le plus court entre deux entités dans un graphe $\mathcal{KG}$ en suivant les liens `is-a` : $dis(a, b) = min_i(N_i)$, N_i longueur de $C_i \in \mathcal{C}$, $\mathcal{C}$ est l'ensemble des chemins entre a et b. Plusieurs améliorations de cette mesure de base ont été proposées pour prendre en compte le profondeur des nœuds dans la hiérarchie (relations hiérarchiques) (Wu & Palmer, 1994; Li *et al.*, 2003; Paul *et al.*, 2016). La plupart de ces mesures se base sur le calcul du LCS qui a montré un intérêt pour des tâches d'extraction d'information du web de données : désambiguïsation et liage d'entités, détection de communautés de données RDF ou extraction automatique de propriétés partagées entre ressources (Colucci *et al.*, 2016).

Mesures basées sur les propriétés (*feature-based*). Ces mesures complètent les méthodes basées sur le chemin en considérant le degré de chevauchement entre les propriétés des entités comparées. La similarité est calculée comme fonction des propriétés en commun et des différences entre les entités. La mesure de base adoptée est celle définie par *Tversky* (Tversky, 1977) : $sim_t(a, b) = \alpha.f(P(a) \cap P(b)) - \beta.f(P(a) \setminus P(b)) - \gamma.f(P(b) \setminus P(a))$, avec $P(a)$ et $P(b)$ respectivement les propriétés des entités a et b. Plusieurs méthodes ont été proposées (Rodriguez & Egenhofer, 2003; Petrakis *et al.*, 2006) selon le choix de la nature des propriétés (e.g. dans WordNet, les *synsets* et *glosses* ont été utilisées) et le calcul des paramètres de pondération α, β et γ.

Mesures basées sur le contenu (*information content*). Ces mesures reposent sur des corpus de textes pour calculer des probabilités sur l'occurrence des mots et des thésaurus (e.g. WordNet) pour le calcul des hyponymes des concepts (Sánchez & Batet, 2011; Traverso-Ribón & Vidal, 2015), un aspect qui sort du cadre de cette étude.

Les mesures de similarité basées purement sur le graphe de connaissances (liens, propriétés) se caractérisent par leur simplicité et efficacité. Elles exploitent le réseau de sommets et de liens étiquetés, contrairement aux méthodes basées sur le contenu qui nécessitent des sources de données externes. Cependant, ces mesures considèrent les aspects des ressources en isolation et représentent moins l'intégralité de l'information autour de ces nœuds. Dans (Traverso *et al.*, 2016), les auteurs proposent une mesure de similarité qui combine différents aspects d'une entité et montrent qu'elle donne une meilleure corrélation avec les valeurs de référence. Ces aspects sont les voisins, la hiérarchie et le degré d'un nœud ou sa spécificité. La définition de cette mesure se rapproche de notre objectif quant à la représentation des différents aspects des ressources dans un graphe. Toutefois, telle qu'elle est définie, elle n'est pas applicable à notre cas comme, par construction du graphe de contexte, l'aspect degré d'un nœud (nombre de liens incidents) n'est pas conservé (notamment pour les nœuds terminaux et appartenant à la *T-Box*). Seuls les deux aspects voisins et hiérarchie peuvent être exploités.

La validation du graphe dans notre approche se repose ainsi sur ces deux aspects en combinant deux types de mesures :

— les mesures basées sur les liens comme elles permettent de couvrir l'aspect structurel dans un graphe (structure locale d'un nœud) ;
— les mesures basées sur les propriétés comme elles exploitent plus de connaissances sémantiques en évaluant à la fois les points communs et les différences.

4.3 Une mesure de similarité pour la validation de graphes de contexte

Nous présentons dans cette section une nouvelle mesure de similarité qui repose sur les liens taxonomiques et les propriétés des entités dans un graphe de connaissances afin de valider l'hypothèse de la section 4.1. Cette mesure se compose de deux parties. La première partie sert à valider la structure du graphe en suivant les liens taxonomique (de type `is-a`) pour comparer deux entités. Nous utilisons pour ceci la mesure de *Wu et Palmer* (Wu & Pal-

mer, 1994).

Définition. Similarité liens. Soient a et b deux entités dans le graphe, N_1 et N_2 respectivement le nombre de liens `is-a` à partir de a et b jusqu'à leur LCS, N_3 est le nombre de liens `is-a` du LCS à la racine du graphe (racine de la A-Box). La similarité $sim_l(a, b)$ entre a et b est calculée comme suit :

$$sim_l(a, b) = \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3} \qquad (1)$$

La deuxième partie est basée sur les propriétés et suit le principe proposé dans le modèle de *Tversky* (décrit dans la section 4.2) qui considère que la similarité entre deux entités est une fonction de leurs propriétés communes et distinctives. Nous considérons en deuxième lieu l'ensemble des couples (propriété,valeur). La même définition est utilisée pour les deux mesures avec l'ensemble des propriétés ou des couples (propriété,valeur).

Définition. Similarité propriétés. Soient a et b deux entités dans le graphe, $\mathcal{P}_a = \{p : (a, v) \mid v \in \mathcal{U} \cup \mathcal{L}\}$ et $\mathcal{P}_b = \{p : (b, v) \mid v \in \mathcal{U} \cup \mathcal{L}\}$ respectivement l'ensemble des propriétés de a et b. La similarité $sim_p(a, b)$ entre a et b est calculée en fonction de la cardinalité de leurs propriétés comme suit :

$$sim_p(a, b) = \frac{|\mathcal{P}_a \cap \mathcal{P}_b|}{|\mathcal{P}_a \setminus \mathcal{P}_b| + |\mathcal{P}_b \setminus \mathcal{P}_a| + |\mathcal{P}_a \cap \mathcal{P}_b|} \qquad (2)$$

Définition. Similarité propriété-valeur. Soient $\mathcal{KG}$ un graphe de connaissances et a et b deux entités dans $\mathcal{KG}$, $\Sigma_a = \{(p, v) \mid p \in \mathcal{P}_a, v \in \mathcal{U} \cup \mathcal{L}\}$ et $\Sigma_b = \{(p, v) \mid p \in \mathcal{P}_b, v \in \mathcal{U} \cup \mathcal{L}\}$ respectivement l'ensemble des couples propriété-valeur de a et b. La similarité $sim_{pv}(a, b)$ entre a et b est calculée en fonction de la cardinalité de l'ensemble des couples comme suit :

$$sim_{pv}(a, b) = \frac{|\Sigma_a \cap \Sigma_b|}{|\Sigma_a \setminus \Sigma_b| + |\Sigma_b \setminus \Sigma_a| + |\Sigma_a \cap \Sigma_b|} \qquad (3)$$

Définition. Mesure de similarité agrégée. Soient $\mathcal{KG}$ un graphe de connaissances et a et b deux entités dans $\mathcal{KG}$, la mesure de similarité agrégée est définie comme suit :

$$sim(a, b) = \top(sim_l(a, b), sim_p(a, b), sim_{pv}(a, b)), \qquad (4)$$

avec $\top$ est la moyenne des similarités précédentes. Toutes les mesures sont normalisées dans l'intervalle $[0, 1]$, où un score égale à 0 signifie que les ressources comparées sont dissemblables, et le score 1 signifie que les ressources sont identiques.

5 Expérimentations et validation sur données de Data&Musée

5.1 Data&Musée

Ce travail est conduit dans le cadre du *projet Data & Musée* [4]. Ce projet vise à améliorer les capacités de différentes institutions culturelles en agrégeant et analysant des données en provenance de ces différentes institutions. Les données récoltées et traitées seront utilisées dans l'objectif de l'élargissement, la fidèlisation et la meilleure compréhension de leurs publics. Début 2019, les institutions partenaires sont les 14 musées de Paris Musées et les 84 monuments du Centre des Monuments Nationaux. Nous présentons dans la suite la constitution d'un graphe de contexte pour ces institutions.

4. Data & Musée, sélectionné dans le 23ième appel à projet du Fonds Unique Interministériel (FUI) et certifié par Cap Digital et Imaginove. `http://datamusee.fr/le-projet/`.

5.2 Création d'un graphe de contexte pour Data&Musée

Comme nous l'avons vu à la section 3, le graphe de contexte est construit à partir d'une liste d'entités, représentées par leurs URIs. La liste d'entité est soit choisie par un expert d'un domaine, soit constituée par des entités évidentes (par exemple, pour les musées de Paris Musées, nous cherchons à la main une entité de DBpedia-fr correspondant à chaque musée).

Le processus d'extraction du graphe de contexte est paramétré par la dimension N du graphe. C'est un processus récursif pour la collecte des nœuds voisins qui dépend du choix de N. Partant des travaux décrits dans la section 2 et en se basant sur les observations faites durant les expérimentations sur nos données, nous considérons que $N = 2$ est un bon choix pour la dimension d'un graphe de contexte.

Le contexte pour Paris Musées a été ainsi construit avec une profondeur de 2. Le cœur du contexte a donc une profondeur de 1. L'étude de l'impact du choix de cette profondeur sort du champs de cet article. Une *blacklist* a été créée comportant essentiellement tous les éléments de la *T-Box*, considérés comme nœuds terminaux. En effet, par exemple, si un nœud nous amenait à `owl:Thing` et qu'on suivait les liens à partir de là, nous ramènerions 1527645 entités pas nécessairement en rapport avec notre domaine.

Le tableau 1 donne une description d'un graphe de contexte extrait de DBpedia-fr pour la profondeur $N = 2$. Nous testons différents réglages pour la constitution d'un tel graphe. Les chiffres dans ce tableau ne sont donc qu'une indication sur un exemple.

	Contexte $\mathcal{CG}$	DBPedia-fr	%
Nœuds distincts	451653	10515624	4,29
Prédicats distincts	2310	20322	11,36
Liens	5150179	185404534	2,78
Liens par nœud (moyenne)	11,4	17,6	

TABLE *1 – Description d'un graphe de contexte $\mathcal{CG}$ extrait de DBPedia-fr avec $N = 2$*

Il est normal qu'il y ait moins de liens par nœuds dans $\mathcal{CG}$ que dans DBPedia-fr, puisque par construction nous avons éliminé certains liens peu porteurs d'information dans notre cadre applicatif comme expliqué plus haut.

Nous avons donc un nombre L de liens 36 fois inférieur et un nombre S de sommets 23 fois inférieur dans le $\mathcal{CG}$ que dans le $\mathcal{KG}$. Sur un algorithme qui est en $O(L + S)$ -comme le parcours en largeur (*Breadth-first search*)-, nous pouvons donc anticiper un gain d'un facteur de l'ordre 30, ce qui peut fortement contribuer à l'applicabilité de certaines méthodes. Les gains peuvent devenir considérables sur des algorithmes tels que ceux de recherche du plus court chemin entre deux nœuds si on souhaite donner un poids aux liens où on peut être en $O(S^2)$.

Nous devrons approfondir ces questions d'apport pour la scalabilité, mais ces premiers indices sont favorables. Nous allons voir dans la suite d'autres indicateurs qui plaident en faveur du graphe de contexte.

5.3 Validation du contexte obtenu

5.3.1 Couverture du domaine par le graphe de contexte

Pour évaluer la pertinence de notre contexte, nous avons voulu voir s'il conserve une bonne couverture des principaux éléments nous concernant dans la base de données Joconde.

La base Joconde est constituée de metadonnées concernant près de 600000 œuvres du patrimoine français et est disponible en Open Data. Chaque œuvre est principalement décrite par le lieu où elle se trouve (ville et institution), son ou ses créateurs, son titre, les techniques dont elle relève et des informations temporelles. Cette base est importante dans notre projet

puisqu'elle apporte un nombre considérable de données qui font autorité pour la description du patrimoine français.

Le tableau 2 illustre la couverture de Joconde par notre graphe de contexte. Nous avons pris les 10 villes associées au plus grand nombre d'œuvres dans la base Joconde et trouvées dans DBPedia-Fr et avons vérifié qu'elles sont bien dans notre graphe de contexte. Nous avons procédé de même pour les créateurs et les musées associés à des œuvres.

	Liste	Dans $\mathcal{CG}$
Villes	Paris, Saint-Germain-en-Laye, Marseille, Strasbourg, Sèvres, Chantilly, Bordeaux, Montauban, Communauté urbaine Creusot-Montceau, Rennes	10/10
Domaines	Dessin, Archéologie, Peinture, Ethnologie, Estampe, Sculpture, Photographie, Céramique, Costume, Néolithique	10/10
Créateurs	A.Rodin, H.Chapu, E.Boudin, G.Moreau, JB.Barla, Y.Jean-Haffen, T.Chassériau, Manufacture nationale de Sèvres, JBC.Corot, E.Delacroix	9/10
Musées	Louvre, Musée d'Archéologie nationale, Cité de la céramique, Musée Rodin, Musée Condé, Musée Ingres, Musée des beaux-arts(Strasbourg), Musée des beaux-arts(Rennes), Musée des beaux-arts(Angers), Musée Gustave-Moreau	10/10

TABLE *2 – Couverture de la base Joconde par notre graphe de contexte*

On voit que la couverture est excellente. Un seul créateur n'a pas été trouvé, *J.B.Barla* : il s'agit d'un naturaliste qui a réalisé de nombreux dessins de plantes, mais n'est pas bien référencé par DBpedia-Fr à ce sujet et donc n'est pas reconnu comme un élément de notre domaine. Pour la couverture en nombre d'œuvres, on a, par exemple, 333114 œuvres associées aux villes mentionnées, soit plus de la moitié des œuvres capturées pour seulement 10 villes.

Cela démontre que notre graphe de contexte assure une bonne couverture de la base Joconde, qui est une des plus importantes pour le domaine du patrimoine culturel français.

5.3.2 Pertinence d'un graphe de contexte par mesure de similarité

Dans cette section nous montrons que des propriétés du graphe d'origine, importantes pour nos travaux, sont préservées dans le graphe de contexte construit.

Similarité liens

D'abord, notons que puisque nous récupérons les types (liens `is-a`) de toutes les entités présentes dans le $\mathcal{CG}$, par construction, le LCS de deux entités calculé sur le $\mathcal{KG}$ (DBpedia-Fr) et sur notre $\mathcal{CG}$ sont identiques pour toutes les entités. Cela constitue un premier indice de pertinence du $\mathcal{CG}$.

Cette première propriété permet d'affirmer que, pour chaque paire d'entités de notre $\mathcal{CG}$, la mesure de similarité de *Wu-Palmer* (formule (1), section 4.3) obtenue sur notre $\mathcal{CG}$ est identique à celle obtenue sur le $\mathcal{KG}$, puisqu'elle ne dépend que

— des mesures de LCS qui sont identiques sur les deux graphes,
— de la distance du LCS à la racine de la *T-Box* utilisée, qui est identique pour les deux graphes puisque nous avons inclu dans notre $\mathcal{CG}$ la *T-Box* de DBPedia-Fr.

Nous pourrons donc utiliser cette mesure de similarité sur notre $\mathcal{CG}$ sans perte d'information. Cela constitue un deuxième indice de pertinence de notre $\mathcal{CG}$.

Similarité propriétés

Nous avons utilisé la mesure de similarité des propriétés des entités définie par la mesure de *Tversky* (formule (2), section 4.3). Comme toutes les propriétés ne sont pas conservées

dans notre $\mathcal{CG}$, cette mesure de similarité donne des résultats différents sur le $\mathcal{CG}$ et sur le $\mathcal{KG}$. Dans ce cas, utiliser une corrélation de rang comme métrique pour évaluer la relation entre deux variables est un bon indicateur de préservation de cette mesure. Nous définissons ainsi une propriété de corrélation de rang et nous avons vérifié sa validité sur $\mathcal{CG}$.

Propriété. Corrélation de rang. Soient a, b et c trois entités de $\mathcal{CG}$ tels que $a, b, c \notin \overline{\mathcal{V}}$. Une corrélation de rang existe entre les paires d'entités (a, b) et (a, c) si :

$$sim_p^{\mathcal{KG}}(a,b) > sim_p^{\mathcal{KG}}(a,c) \Rightarrow sim_p^{\mathcal{CG}}(a,b) > sim_p^{\mathcal{CG}}(a,c) \quad (5)$$

avec $sim_p^{\mathcal{KG}}$ et $sim_p^{\mathcal{CG}}$ sont les mesures de similarité respectives sur $\mathcal{KG}$ et $\mathcal{CG}$.

Pour vérifier la propriété (5), nous avons appliqué l'algorithme suivant sur un ensemble d'entités a et e choisies aléatoirement parmi l'ensemble des nœuds centraux :

— choisir $\{a_1, ..a_m\} \notin \overline{\mathcal{V}}$ et $\{e_1, ..e_n\} \notin \overline{\mathcal{V}}$

— $\forall l \in \{1, 2, ..m\}, \forall i \in \{1, 2, ..n\}$, calculer $sim_p^{\mathcal{KG}}(a_l, e_i)$ et $sim_p^{\mathcal{CG}}(a_l, e_i)$

— $\forall (e_i, e_j) \mid i, j \in \{1, 2, ..n\}, i \neq j$, vérifier la condition :

$$sim_p^{\mathcal{KG}}(a_l, e_i) > sim_p^{\mathcal{KG}}(a_l, e_j) \Rightarrow sim_p^{\mathcal{CG}}(a_l, e_i) > sim_p^{\mathcal{CG}}(a_l, e_j)$$

et compter le nombre de fois où elle est vérifiée.

Nous avons choisi $m = 100$ et $n = 10$ puis $m = 20$ et $n = 20$ et nous avons effectué $\frac{n(n+1)}{2} \times m$ vérifications pour valider la mesure de corrélation de rang. Le tableau 3 montre les résultats d'une série de tests pour ces différentes valeurs de m et n :

m	n	Nb. vérifications	Nb. succès	%
100	10	5500	5070	92,18
100	10	5500	5137	93,40
20	20	4200	3778	89,95
20	20	4200	3873	92,21

TABLE *3 – Corrélation de rang sur $\mathcal{KG}$ et $\mathcal{CG}$*

Ainsi, lorsque nous utilisons cette similarité pour comparer des éléments et en proposer à un utilisateur, dans 90% des cas ou plus, la proposition que nous pourrons faire sera identique à celle que nous aurions faite sur le $\mathcal{KG}$ complet.

Coefficient de corrélation rang-ordre de *Spearman*. Nous avons également calculé le coefficient de corrélation de *Spearman* qui est la métrique classique utilisée dans la littérature pour évaluer les mesures de similarité. Cette corrélation évalue la relation monotone entre deux variables.

De la même manière, nous avons calculé ce coefficient pour les valeurs $sim_p^{\mathcal{KG}}(a, e)$ et $sim_p^{\mathcal{CG}}(a, e)$ sur un ensemble d'entités a et e choisies aléatoirement. Nous avons réitéré ce processus plusieurs fois et nous avons calculé la mesure globale $sim^{\mathcal{KG}}(a, e)$ et $sim^{\mathcal{CG}}(a, e)$. Les résultats, décrits dans le tableau 4, montrent de très bonnes valeurs de corrélation.

La similarité globale $sim(a, e)$ est calculée comme moyenne de $sim_l(a, e)$ et $sim_p(a, e)$. Les expérimentations sur la similarité $sim_{pv}(a, e)$ définie par la formule (3) donnent de moins bon résultats. Ceci est dû au fait que toutes les propriétés-valeurs ne sont pas conservées dans $\mathcal{CG}$ (comme décrit plus haut). Les premiers résultats nous paraissent très encourageants et nous incitent à poursuivre l'exploitation de graphes de contexte dans le projet Data&Musée.

IC 2019

m	n	Corrélation Spearman
20	10	0.944
20	20	0.949
20	10	0.964
20	20	0.934

TABLE *4 – Corrélation de Spearman pour mesure de similarité sur $\mathcal{KG}$ et $\mathcal{CG}$*

6 Expérimentations et validation sur données de la base KORE

Dans un cadre général, pour évaluer la similarité entre entités, des données de référence ($benchmark$) existent. Afin de comparer l'utilisation des graphes de contexte $\mathcal{CG}$ avec l'utilisation du graphe $\mathcal{KG}$, nous utilisons le jeu de données de référence KORE (Hoffart *et al.*, 2012). Il contient 21 entités principales dans 5 domaines différents : *IT companies, Hollywood celebrities, Television series, Video games* et *Chuck Norris*. Pour chacune des entités principales, il contient 20 entités classées par similarité par rapport à celle-ci, la plus similaire étant classée en premier. Ceci résulte en 420 paires d'entités classées du plus au moins similaire. Nous utilisons la corrélation de Spearman comme métrique d'évaluation.

Nous avons identifié semi-automatiquement l'ensemble des entités de KORE dans DBpedia. Pour chacun des 5 domaines de KORE nous avons créé un graphe de contexte utilisant comme germes l'ensemble de ses entités qu'on passe en entrée à l'algorithme `CONTEXT`. En sortie nous avons 5 graphes de contextes sur lesquels nous évaluons la similarité pour les paires d'entités du jeu de données. Nous avons effectué les calculs de similarité entre les entités de KORE sur ces graphes et sur DBPedia, afin de comparer les résultats obtenus.

Le tableau 5 donne les résultats de la corrélation de Spearman entre $\mathcal{KG}$ et $\mathcal{CG}$ sur le jeu de données de référence. Chaque ligne du tableau décrit les valeurs de corrélation pour la mesure de similarité correspondante. La dernière ligne correspond à la corrélation sur le classement de la mesure de similarité calculée comme moyenne des trois précédentes. La colonne 'Moyenne' décrit les valeurs de corrélation pour toutes les entités des domaines de KORE considérés (le traitement du domaine *Video Games* a du être différé pour raison technique).

Mesure	IT Companies	Hollywood Celebrities	Television Series	Chuck Norris	Moyenne
$sim_l(a,b)$	1.0	0.999	0.995	0.898	0.973
$sim_p(a,b)$	0.997	0.998	0.994	0.998	0.997
$sim_{pv}(a,b)$	0.590	0.807	0.646	0.806	0.712
$sim(a,b)$	0.994	0.996	0.957	0.986	0,983

TABLE *5 – Corrélation de Spearman pour mesures de similarité sur $\mathcal{KG}$ et $\mathcal{CG}$(KORE)*

Nous observons sur le tableau que les mesures $sim_l(a,b)$, $sim_p(a,b)$ et $sim(a,b)$ donnent de très bonnes valeurs de corrélation entre les classements obtenus sur $\mathcal{KG}$ et ceux sur $\mathcal{CG}$, ce qui est un élément de confirmation de plus en faveur de l'utilisation de graphes de contexte. Comme dans le cas des graphes de contexte de Paris Musées (section 5.3.2), sur les graphes de contexte de KORE la mesure $sim_{pv}(a,b)$ donne de moins bons résultats. Des tests sont en cours pour améliorer les résultats de cette mesure.

7 Conclusion et perspectives

Dans cet article nous avons présenté la notion de graphe de contexte pour un domaine. Nous le définissons comme un extrait d'un plus grand graphe et qui cible les connaissances sur ce domaine. Nous avons montré que ce graphe peut être construit simplement en partant de quelques entités importantes du domaine utilisant un jeu de données de départ ou avec peu de connaissances expertes si elles sont disponibles. Nous avons aussi montré que le graphe obtenu présente des caractéristiques qui permettent de le substituer au grand graphe pour des exploitations classiques du graphe de connaissance (étude sur la similarité entre des éléments). Dans un proche avenir, nous comptons appliquer cette technique à d'autres domaines et exploiter les graphes de contexte obtenu pour leur appliquer des techniques d'apprentissage sur les graphes.

Références

AL-GHOSSEIN M., ABDESSALEM T. & BARRÉ A. (2018). Open data in the hotel industry : leveraging forthcoming events for hotel recommendation. *J. of IT & Tourism*, **20**(1-4), 191–216.

BEEK W., SCHLOBACH S. & VAN HARMELEN F. (2016). A contextualised semantics for owl :sameas. In *Proceedings of the 13th European Semantic Web Conference*, volume 9678 of *LNCS*, p. 405–419 : Springer.

BENEDETTI F., BENEVENTANO D., BERGAMASCHI S. & SIMONINI G. (2019). Computing inter-document similarity with context semantic analysis. *Information Systems*, **80**, 136 – 147.

BHATT S., PADHEE S., SHETH A., CHEN K., SHALIN V., DORAN D. & MINNERY B. (2019). Knowledge graph enhanced community detection and characterization. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, p. 51–59.

CHENIKI N., BELKHIR A., SAM Y. & MESSAI N. (2016). Lods : A linked open data based similarity measure. In *2016 IEEE 25th International Conference on Enabling Technologies : Infrastructure for Collaborative Enterprises (WETICE)*, p. 229–234.

COLUCCI S., DONINI F. M., GIANNINI S. & SCIASCIO E. D. (2016). Defining and computing least common subsumers in rdf. *J. Web Semant.*, **39**, 62–80.

CONSOLI S., MONGIOVÌ M., NUZZOLESE A. G., PERONI S., PRESUTTI V., RECUPERO D. R. & SPAMPINATO D. (2015). A smart city data model based on semantics best practice and principles. In *WWW'15*.

GYRARD A., ATEMEZING G., BONNET C., BOUDAOUD K. & SERRANO M. (2016). Reusing and unifying background knowledge for internet of things with lov4iot. In *2016 IEEE 4th International Conference on Future Internet of Things and Cloud (FiCloud)*, p. 262–269.

HARISPE S., SÁNCHEZ D., RANWEZ S., JANAQI S. & MONTMAIN J. (2014). A framework for unifying ontology-based semantic similarity measures : A study in the biomedical domain. *Journal of Biomedical Informatics*, **48**, 38 – 53.

HOFFART J., SEUFERT S., BA NGUYEN D., THEOBALD M. & WEIKUM G. (2012). Kore : Keyphrase overlap relatedness for entity disambiguation.

HULPUS I., HAYES C., KARNSTEDT M. & GREENE D. (2013). Unsupervised graph-based topic labelling using dbpedia. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, p. 465–474.

LI Y., BANDAR Z. A. & MCLEAN D. (2003). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, **15**(4), 871–882.

LODI G., ASPRINO L., NUZZOLESE A. G., PRESUTTI V., GANGEMI A., RECUPERO D. R., VENINATA C. & ORSINI A. (2017). *Semantic Web for Cultural Heritage Valorisation*, In *Data Analytics in Digital Humanities*, p. 3–37. Springer International Publishing : Cham.

LUO Y., WANG Q., WANG B. & GUO L. (2015). Context-dependent knowledge graph embedding. p. 1656–1661.

ORAMAS S., OSTUNI V., DI NOIA T., SERRA X. & DI SCIASCIO E. (2016). Sound and music recommendation with knowledge graphs. *ACM Transactions on Intelligent Systems and Technology (TIST)*, **8**, 1–21.

PAUL C., RETTINGER A., MOGADALA A., KNOBLOCK C. A. & SZEKELY P. A. (2016). Efficient graph-based document similarity. In *International Semantic Web Conference (ISWC) 2016*.

IC 2019

PETRAKIS E. G. M., VARELAS G., HLIAOUTAKIS A. & RAFTOPOULOU P. (2006). X-similarity : Computing semantic similarity between concepts from different ontologies. *Journal of Digital Information Management (JDIM*, **4**.

RAAD J., PERNELLE N. & SAÏS F. (2017). Détection de liens d'identité contextuels dans une base de connaissances. In *IC 2017 - 28es Journées francophones d'Ingénierie des Connaissances*, p. 56–67, Caen, France.

RADA R., MILI H., BICKNELL E. & BLETTNER M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, p. 17–30.

RODRIGUEZ M. A. & EGENHOFER M. J. (2003). Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering*, **15**(2), 442–456.

SÁNCHEZ D., BATET M., ISERN D. & VALLS A. (2012). Ontology-based semantic similarity : A new feature-based approach. *Expert Syst. Appl.*, **39**, 7718–7728.

SHEN W., WANG J. & HAN J. (2015). Entity linking with a knowledge base : Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, **27**(2), 443–460.

SHI J., GAO H., QI G. & ZHOU Z. (2017). Knowledge graph embedding with triple context. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, p. 2299–2302.

SOUALAH ALILA F., COUSTATY M., REMPULSKI N. & DOUCET A. (2016). Datatourism : designing an architecture to process tourism data. In *IFITT and ENTER 2016 Conferences*.

SÁNCHEZ D. & BATET M. (2011). Semantic similarity estimation in the biomedical domain : An ontology-based information-theoretic perspective. *Journal of Biomedical Informatics*, **44**(5), 749 – 759.

TRAVERSO I., VIDAL M.-E., KÄMPGEN B. & SURE-VETTER Y. (2016). Gades : A graph-based semantic similarity measure. In *Proceedings of the 12th International Conference on Semantic Systems*, SEMANTiCS 2016, p. 101–104.

TRAVERSO-RIBÓN I. & VIDAL M. (2015). Exploiting information content and semantics to accurately compute similarity of go-based annotated entities. In *2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, p. 1–8.

TVERSKY A. (1977). Features of similarity. *Psychological Review*, **84**(4), 327–352.

WANNOUS R., MALKI J., BOUJU A. & VINCENT C. (2017). Trajectory ontology inference considering domain and temporal dimensions—application to marine mammals. *Future Generation Computer Systems*, **68**, 491 – 499.

WU Z. & PALMER M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, p. 133–138.

Apports des bases de connaissances RDF à la détection de vérité et vice versa

Valentina Beretta[1], Sébastien Harispe[2], Sylvie Ranwez[2], and Isabelle Mougenot[3]

[1] MIDN, IRD, Marseille, France

[2] LGI2P, IMT Mines Ales, Univ Montpellier, Ales, France

[3] UMR 228 Espace Dev UM, Maison de la Télédétection, Montpellier, France

Résumé : La détection de vérité sur le Web (ou Truth Discovery), est un domaine de recherche qui a émergé ces dernières années pour contrer les dangers de la désinformation. Elle vise à identifier les faits (assertions *vraies*) lorsque des assertions contradictoires sont émises par différentes sources. Partant de l'hypothèse que les assertions vraies sont fournies par des sources fiables, et que les sources fiables fournissent des assertions vraies ; les modèles de recherche de vérité de la littérature calculent itérativement la confiance dans une assertion et la fiabilité des sources. Il est ainsi possible d'identifier les assertions qui se révèlent vraies.

Dans l'étude presentée dans Beretta *et al.* (2018), nous avons exploité les connaissances du domaine exprimées sous la forme de bases de connaissances RDF, pour améliorer les performances des méthodes actuelles de recherche de vérité. Notre postulat est que la connaissance *a priori* d'un domaine peut conforter certaines assertions et donc influencer le processus de détection de vérité. En étudiant les cooccurrences entre les faits présents dans une base de connaissances RDF, il est possible d'identifier des motifs sous forme de règles qui renforcent la confiance accordée à certaines affirmations exprimées sous la forme de triplets <sujets, prédicat, valeur>. Par exemple, une analyse de la base de connaissances peut permettre de déduire que la majorité des personnes qui parlent espagnol, sont nées en Espagne. Ce constat peut être pris en compte dans un processus de recherche de vérité concernant le lieu de naissance du peintre Pablo Picasso. Si on observe que Pablo Picasso parle couramment espagnol, la confiance attribuée à l'assertion <Pablo Picasso, est né en, Espagne> doit être renforcée. Il en va de même pour les assertions qui contiennent des valeurs plus génériques, à l'exemple de <Pablo Picasso, est né en, Europe>.

Dans notre approche, l'identification des règles est réalisée à l'aide d'AMIE+ (Galárraga *et al.*, 2015). A partir de ces règles, nous avons calculé un coefficient propulseur (ou booster) qui traduit les cooccurrences récurrentes entre différents faits sous la forme de mesures de qualité, afin de renforcer la confiance dans certaines valeurs (et donc dans certaines assertions) lors du processus de détection de vérité. Ce coefficient propulseur représente le degré de soutien (ou caution) conféré à une assertion en exploitant le contenu de la base de connaissances. Ainsi le postulat de base du processus de recherche de vérité présenté en introduction en sera modifié et nous considérerons désormais que les faits (vérités) sont des assertions proposées par des sources fiables et/ou qui sont renforcées par un coefficient booster élevé, en considération de règles d'association extraites de la base de connaissances. 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 assertion sera donc paramétrable afin d'accorder plus d'importance soit à la fiabilité des sources, soit à l'information contenue dans la base de connaissances en fonction du contexte et/ou de la qualité de la base.

Une évaluation de l'approche a été réalisée à l'aide de jeux de données réels. L'objectif principal était d'évaluer l'impact de la prise en compte des connaissances *a priori* lors de l'utilisation des modèles de recherche de vérité dans un scénario réel, doté de spécificités propres. Nous avons adapté le modèle de découverte de vérité Sums, défini dans (Pasternack & Roth, 2010). La méthode Sums adopte une procédure itérative dans laquelle le calcul de la fiabilité associée à une source et le calcul de la confiance associée à une assertion sont alternés jusqu'à atteindre une convergence. Nous l'avons adaptée en modifiant le calcul de la confiance d'une assertion pour qu'il prenne en compte le coefficient propulser. Nous montrons notamment que les modifications apportées au modèle Sums permettent d'obtenir des gains de performance de l'ordre de 18% par rapport à l'approche Sums seule. Par comparaison avec l'existant, ce gain de performances permet de placer notre modèle parmi les les modèles les plus performants du domaine du recherche de vérité.

Références

BERETTA V., HARISPE S., RANWEZ S. & MOUGENOT I. (2018). Combining truth discovery and RDF knowledge bases to their mutual advantage. In *The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8-12, 2018, Proceedings, Part I*, p. 652–668.

IC 2018

GALÁRRAGA 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.
PASTERNACK J. & ROTH D. (2010). Knowing what to believe (when you already know something). In *Proceedings of the 23rd International Conference on Computational Linguistics*, p. 877–885.

Meeting Intents Detection Based on Ontology for Automatic Email Answering

Manon Cassier[1], Zied Sellami[2] and Jean-Pierre Lorré[2]

[1] Laboratoire AGORA, Université de Cergy-Pontoise
manon.cassier@u-cergy.fr

[2] Linagora GSO, 75 route de Revel, 31500 Toulouse, France
{zsellami,jplorre}@linagora.com

Abstract : Automatic email answering is a difficult AI problem that combines classification, natural language understanding and text generation techniques. We present an original approach and a tool based on an ontology to automatically reply to meeting emails. We constructed the ontology from a French corpus of 1150 emails in which the concepts represent detailed meeting intents (proposing a meeting, cancelling a meeting, rescheduling a meeting) and different answer templates. Each intent concept is a semantic rule formalized according to the FrameNet methodology. These rules are used to detect intents in emails and also to extract relevant information (such as date, time or person) used for generating replies. The main advantage of our approach is the generation of more precise answers than those proposed by other approaches. We tested the intent detection step on a set of 297 emails and compared it with different supervised machine learning algorithms. Obtained results are encouraging, with an accuracy 20% higher than results obtained with other algorithms.

Mots-clés : Ontology engineering, knowledge acquisition from text, knowledge-based recommendation systems.

1 Introduction

Automatic email answering is an interesting feature in a business context. According to the Radicati Group 2017 report[1] an employee receives in a day 88 emails and sends 34. Email management would represent between 5 and 10 hours of an employee's time over a month. A too massive use of emails can heavily impact work productivity by causing information overload. Automatic reply to emails therefore represents a considerable challenge.

Our work fits in a larger project to design an Open-Source collaborative platform for businesses called OpenPaaS[2] including a mailbox, an enterprise social network and a shared agenda. Considering the problem of email management, our goal is to propose an email processing assistant that would be able to assist the user by automatically prioritizing emails, sending notifications when urgent emails are received and generating answers.

In this paper, we present our approach for automatic email answering based on an ontology to automatically detect meeting intents in emails and to generate appropriate answers using text pattern matching and named entities detection.

The paper is organized as follows. In Section 2, we review related work. In Section 3, we detail our approach by describing the ontology and how it is used for detecting meeting intents and for generating answers. In Section 4, we compare our system to other intents detection techniques and we test the quality of the generated answers. In Section 5 we discuss the limits and possible improvements of our approach. Section 6 concludes.

[1]Report available at https://www.radicati.com/wp/wp-content/uploads/2017/01/Email-Statistics-Report-2017-2021-Executive-Summary.pdf

[2]OpenPaaS: https://open-paas.org/

2 Related Work

The problem of automatic emails answering has been studied frequently in the last ten years. (Katakis *et al.*, 2006) offer a complete view of the different approaches frequently used for email classification, automatic email summary generation or email answering. Automatically answering to an email requires two steps: detection of intents to find those that require an answer and the generation of the answer according to the detected intent.

These steps were mainly studied through different approaches, exploiting either machine learning algorithms or text-pattern matching.

(Malik *et al.*, 2007) use key phrase extraction and text similarity to perform email classification. A Naïve Bayes model automatically detects key phrases of length up to 3 words in the incoming email and maps them with question-answers pairs already identified to choose which one is similar.

(Kannan *et al.*, 2016) propose Smart Reply for Inbox by Gmail (Google) by using recurrent neural networks (RNNs) as long short-term memory (LSTM) networks to predict the most likely responses for an incoming message. Most of the answers proposed by this system are appropriate for informal context – for instance for yes/no questions. However, these answers do not correspond to the precise response model we expect for professional use. Moreover, these types of approaches require access to a considerable body of data that we do not have.

We believe that symbolic approaches can describe more precisely the elements that the answers must contain. (Carvalho, 2008) introduces the notion of Email acts inspired by the Speech Act Theory of (Austin, 1962) and (Searle, 1969). These acts are described as "*noun-verb pairs that express typical intentions in email communication – for instance, to request for information, to commit to perform a task or to propose a meeting*". (Carvalho, 2008) proposes a taxonomy of these "Email Speech Acts" (i.e. "intents" or "intentions") that are associated with some verbs or nouns. This taxonomy contains only general concepts[3], which are not enough to describe the phenomena that we wish to process in emails. In our approach, we focus on the *meet* act which may itself be divided into several sub-intents.

(Sneiders, 2010) and (Kosseim *et al.*, 2001) propose two question-answering approaches based on symbolic rules to answer emails. These approaches use pattern matching to find specific questions in emails. For each question, a set of standard answers are defined and used on generating a draft reply. In our work, we want to detect answers, affirmations, notifications, assertions, and so on. We want to model a more complex structure than regular expressions. We use FrameNet[4] to formalize semantic frames to detect intents on emails.

The FrameNet project (Ruppenhofer *et al.*, 2016) gives a formalization of semantic frames as described in Charles J. Fillmore's works (Fillmore, 1976). Each semantic frame is represented by one or more frame elements (FEs) that are evocated by words called lexical units (LUs). These frames are only checked if the frame elements are actually present in a sentence.

For example, in case of proposing an appointment by email, the semantic frame *appointment proposal* in the sentence "Je vous propose de faire une réunion jeudi prochain" (I suggest we meet next Thursday) can be detected thanks to the lexical units "propose" and "réunion" and the frame elements SPEAKER (i.e. "Je"), PROPOSITION (i.e. "propose") and TIME (i.e. "jeudi prochain").

Our contribution consists on building a domain ontology from a French corpus of corporate emails based on FrameNet formalization. This ontology is used to extract meeting intents from emails and for automatic emails answering.

[3]The taxonomy contains the 7 concepts *request, propose, deliver, commit, directive, commissive* and *meet*.

[4]The FrameNet Homepage: `https://framenet.icsi.berkeley.edu/fndrupal/` (last accessed 2018/06/08.

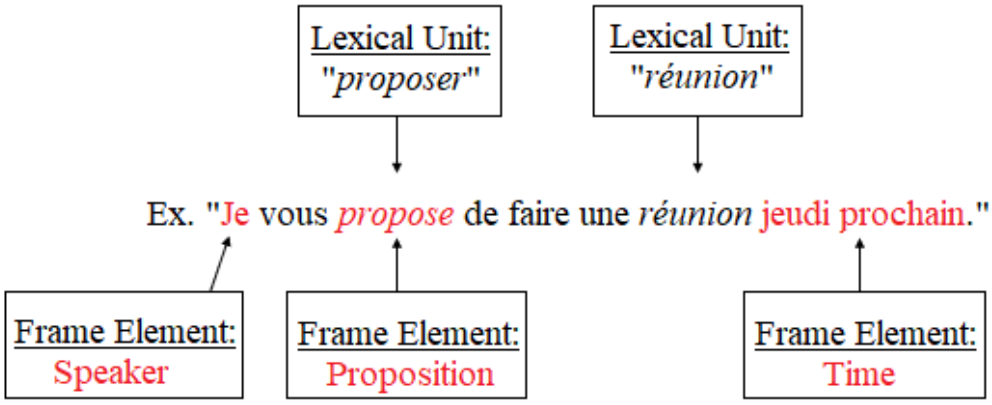


Figure 1: A Frame representation of an appointment proposal.

3 Ontology-based Email Answering

We manually built an OWL ontology[5] from a large corpus of 30000 business emails[6] – empty, English or social network notification emails were removed. The OWL file of the ontology is available on the project github repository[7].

We used the TXM concordancer[8] to explore emails and quickly retrieve those containing meeting intents. This step also allowed us to extract two sets of respectively 1150 annotated emails for the ontology building (with 458 emails containing at least a meeting intent and 692 without any meeting intent) and 177 annotated emails (with 143 emails containing meeting intents) for a subsequent evaluation step.

The first level of the ontology contains 3 core concepts: the Core Intent that models meeting intents, the Core Answer concept that models answer patterns for intents and the Frame Element Concept that models frame elements classes (e.g Time, Date, Address, Person).

3.1 Meeting Intent Concepts

We identified 18 specific meeting intent concepts organized on 3 main concepts (Request, Proposal and Notification) as presented in table 1.

Each intent concept contains lexical units and frame elements (which are Annotation Properties in the ontology) that allow detecting a specific intent in an email. A lexical unit is defined with a String value (lemma form), a part-of-speech tag (e.g. VERB, NOUN, ADJ) or a regular expression (EXPR) and an Annotation Value (Mandatory or Optional). A frame element is defined with a Frame Element Class resource (these are generally what we detect as named entities) or a Frame Element Individual resource – which are defined by a regular expression containing words, expressions or characters that evocate them – and an Annotation Value (Mandatory or Optional). For instance, the *request* concept contains the frame element INTERROGATION – represented by a regular expression containing the strings "?", "est-ce que", "quel", etc. (i.e. the English WH questions words equivalents) – and the lexical units "demander" (ask for), "est-ce possible" (is it possible), "(je|nous) souhait(e|ons)" (I|we wish), etc. The Annotation Value determines if a lexical unit or a frame element is necessary

[5]We build the ontology with the Protégé tool (Stanford Center for Biomedical Informatics Research, 2016): `http://protege.stanford.edu/`.

[6]The corpus can not be distributed for confidentiality reasons.

[7]`https://github.com/openpaas-ng/automatic-email-answering/blob/master/intent6.owl`

[8]TXM Concordancer (Heiden, 2010): `http://textometrie.ens-lyon.fr/`.

or not for identifying the intent on the email (i.e. if the element is specific to the current intent or not). This information helps to compute a score when using the ontology to automatically annotate emails.

Table 1: Description of the core intents of the meet act ontology.

Request	The *request* concept includes both requests for information and request addressed to the recipient to perform some activity related to an appointment. It contains 7 intents: *request an appointment cancellation, request an appointment confirmation, request to schedule an appointment, request an appointment change, request details about an appointment, request availabilities for an appointment* and *request a participation to an appointment*.
Proposal	The *propose* concept includes proposals addressed to the recipient to do something or to take part in something related to an appointment. It contains 3 intents: *propose an appointment, propose an appointment cancellation* and *propose an appointment change*.
Notification	The *notification* concept includes all the messages that just observe a fact about an appointment. It contains 8 intents: *availability confirmation, appointment confirmation, appointment cancellation notification, unavailability notification, availability notification, appointment reminder notification, precisions about an appointment* and *appointment change notification*.

Shared frame elements and lexical units are described in the 3 main concepts so that subconcepts that represent intents inherit them. For example, the frame element Interrogation is transmitted to all intents related to the main concept request differentiating them from intents related to other concepts propose and notification.

A higher concept called meeting containing only lexical units that refer to an appointment (e.g. “entretien”, “RDV”, “reunion”) is used to filter the incoming emails and detect only those containing sentences related to an appointment.

The ontology is enriched by non-hierarchical relations between intents (called implication relation) that model associations of intentions within a same email. For example, in the sentence “Je ne suis pas disponible demain, pouvons-nous décaler l’entretien ?” (I’m not available tomorrow, can we change the time of my interview), unavailability notification intent implies an appointment change request.

3.2 Answer Pattern Concepts

For each meeting intent, we modeled in the ontology a set of answer templates (patterns). These templates were determined by analyzing the responses that people give to each type of appointment email. An answer can be shared between multiple intents.

For example, a template for *accepting meeting proposal* is: « D’accord pour {[DET;values=le,la,l’;default=le] [Appointment_Intent;default=rendez-vous]} {[Time]} {[DET;values=avec;default=avec] [Person;default=]}. ». In this template, Time, Person and Appointment_Intent will be changed with corresponding values recovered from the sentence in the email that instantiated the *accepting meeting proposal* intent.

Note that our system will be able to use the calendar module to check when the person is available for an appointment. This information will be also used to generate a more precise answer.

3.3 The System

Our system is integrated to the OpenPaaS UnifiedInbox module (Figure 2) that communicates with our answers suggestion service through a REST API. The system is open-source and

available on github[9].

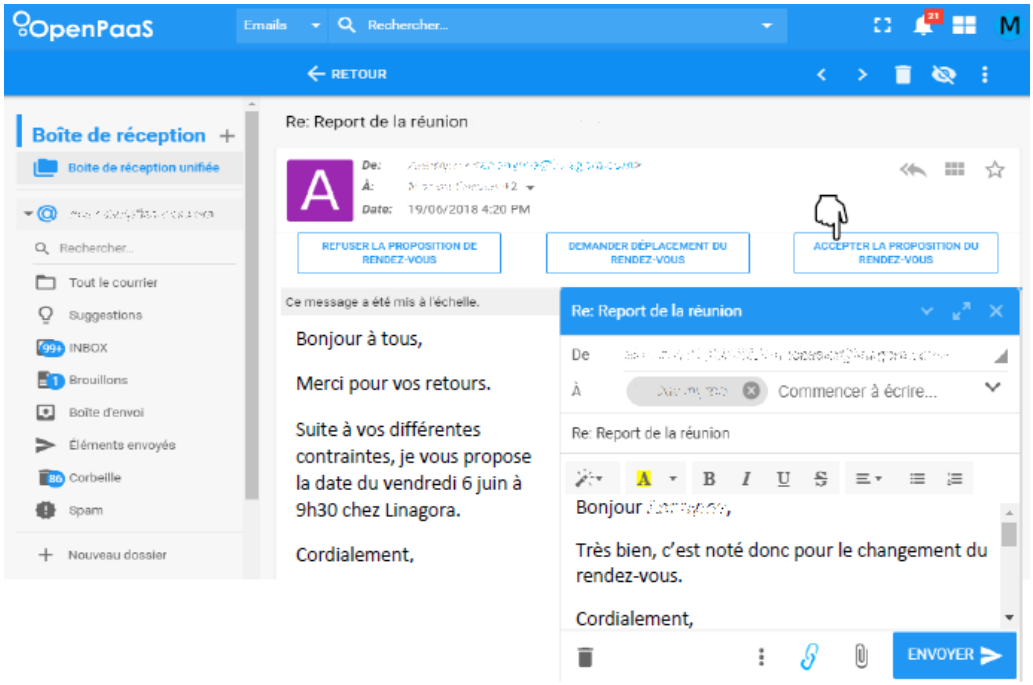


Figure 2: Screenshot of an example of answers suggestions in OpenPaaS.

For each new email, a JSON request is sent to the Web Service module with the text of the email and other metadata (sender, recipients, date). The system analyzes the query and returns a JSON response with answer suggestions as presented in Figure 3.

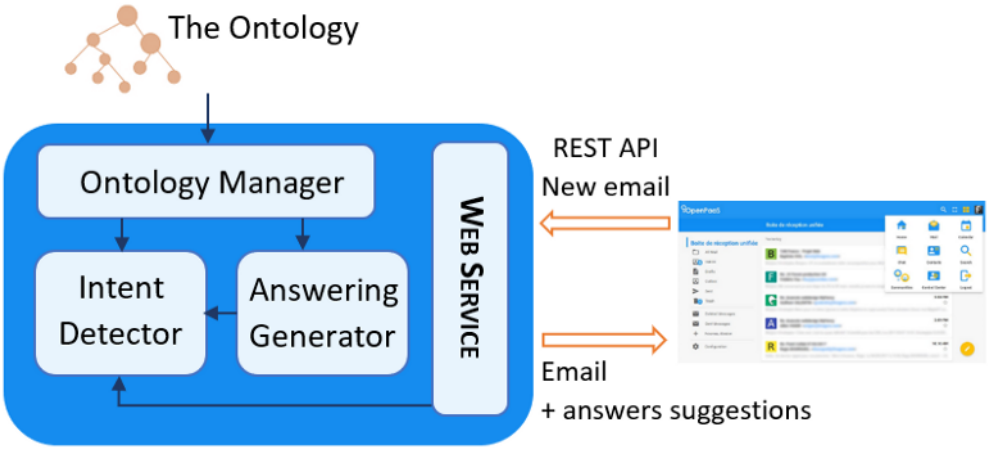


Figure 3: General architecture of the automatic email answering system.

Our system has four main modules:

1. An **Ontology Manager** that uses the Apache Jena API to parse and load the ontology in a Java object.

[9] https://github.com/openpaas-ng/automatic-email-answering

2. An **Intent Detector** that integrates the CoreNLP[10] (Manning *et al.*, 2014) tool which makes tokenization, POS tagging and named entities detection on the text of the email. It exploits at the same time the Duckling Facebook[11] tool that is more successful for the detection of encrypted named entities. When two different named entities are proposed, a resolver automatically selects the proposal made by Duckling Facebook. The text of the email and the lexical units contained in the ontology are then completely stemmed[12] by the Snowball Stemmer[13] to allow easier matching of forms with text shapes. The module retrieves the information available in the ontology and makes a projection of the lexical units of the higher intent meeting on the text of the email to check if the current email contains meeting intents or not. This projection is made on the entire message thread to which the target email belongs to avoid missing emails that do not contain the word "rendez-vous" (appointment) or a synonym but do contain an intention. For example, it allows conserving the sentence "Oui, je suis disponible" (Yes, I'm available) answering to "Êtes-vous disponible demain pour un rendez-vous ?" (Are you available tomorrow for a meeting?). If no intent is detected, the system does not propose any intent and the Web Service returns the message meeting intent not detected. Otherwise, it makes a projection of the other lexical units and frame elements only on the sentences of the incoming email (without taking the message thread or the order of appearance of LUs and FEs in account) and computes a score by assigning a higher weight to units marked as mandatory in the ontology to select the three more likely intents – after several experiments, we also chose to give a higher score to lexical units than frame elements. Finally, the module computes an annotation score (described below) to each intent detected thanks to the matching step.
3. An **Answering Generator** that generates an answer for each intent detected. This module integrates an email generator that recovers the answers templates associated to detected intents. It adds a greeting according to the time – for example, "Bonsoir" (Good evening) after 8:00 pm – and a closing formula as "Cordialement" (Sincerely). It computes a score for each answer, corresponding to the sum of the instances scores of each intention in the email. This module will also exploit the user's calendar to check its availabilities and complete the Time, Person and Appointment_Intent tags in the answers templates.
4. A **Web Service** that returns the three first associated answers in a JSON format.

The annotation score is computed after a filtering of the matching results. The system only keeps concepts of the ontology lowest level (i.e. the intents) that match at least two instances of lexical units or frame elements and of which at least one of the lexical units matched is mandatory.

The score assigned to an intention corresponds to its relevance value (1) in relation to the sentence studied. It corresponds to the sum of its annotation value (2) and its specificity value (3) as described below.

$$Relevance\ value = Annotation\ value + Specificity\ value \tag{1}$$

$$\begin{aligned} Annotation\ value = {} & nbInstances + nbSpecificFE \\ & + 4 * nbSpecificLU \\ & + 4 * nbSpecificMandatoryInstances \\ & + nbMandatoryInstancesOfSuperConcept \end{aligned} \tag{2}$$

[10]CoreNLP: `https://stanfordnlp.github.io/CoreNLP/`

[11]Duckling Facebook: `https://duckling.wit.ai/`

[12]The choice of a stemmer is motivated by the fact that we did not find an open source lemmatizer effective enough for French.

[13]`SnowballStemmer:https://snowballstem.org/`

$$Specificity\ value = Ratio\ 1 + Ratio\ 2 + Ratio\ 3 + Ratio\ 4 \tag{3}$$

$$Ratio\ 1 = \begin{pmatrix} \sum[\frac{1}{nbMandatoryLU}]\ if\ the\ mandatory\ LU\ is\ instanciated \\ \sum[\frac{-1}{nbMandatoryLU}]\ if\ the\ mandatory\ LU\ is\ not\ instanciated \end{pmatrix} \tag{4}$$

$$Ratio\ 2 = \begin{pmatrix} \sum[\frac{1}{nbMandatoryFE}]\ if\ the\ mandatory\ FE\ is\ instanciated \\ \sum[\frac{-1}{nbMandatoryFE}]\ if\ the\ mandatory\ FE\ is\ not\ instanciated \end{pmatrix} \tag{5}$$

$$Ratio\ 3 = \sum[\frac{1}{nbOptionalLU}]\ if\ an\ optional\ LU\ is\ instanciated \tag{6}$$

$$Ratio\ 4 = \sum[\frac{1}{nbOptionalFE}]\ if\ an\ optional\ FE\ is\ instanciated \tag{7}$$

The parameters of the annotation value are obtained empirically. An instance of an intent is correct if it is instantiated with mandatory and specific lexical units or frame elements.

The specificity value is obtained by calculating the ratio of frame elements and lexical units. This formula makes it possible to estimate the level of coverage of the instance in relation to the properties of its class.

We detail next how these scores are calculated for an intent of *request availabilities for an appointment* in the sentence "Pouvez-vous m'indiquer vos disponibilités pour planifier un rendez-vous?" (Please let met know when you're available). The different parameters used to compute the annotation value of the sentence are:

- Number of Instances = 4 ("pouvez-vous", "disponibilités", "planifier" and "rendez-vous")
- Specific FE = 0
- Specific LU = 2 ("disponibilités" and "planifier")
- Specific mandatory instances = 1 ("disponibilités")
- Inherited mandatory instances = 2 ("pouvez-vous" (*request* intent) and "rendez-vous" (*meeting* intent))

$$Annotation\ value = 4 + 4\ *\ (2 + 1) + 2 = 18$$

The *request availabilities for an appointment* contains 14 mandatory LU, 4 optional LU, 1 mandatory FE and 1 optional FE. We calculate the specificity value with the results of the 4 ratios.

$$Ratio\ 1 = \frac{1}{14} + 13 * \frac{-1}{14} \quad Ratio\ 2 = \frac{-1}{1} = -1 \quad Ratio\ 3 = \frac{1}{4} = 0.25 \quad Ratio\ 4 = 0$$

The specificity value of the intent *request availabilities for an appointment* for the sentence is:

$$Specificity\ value = \text{-0.86 - 1 + 0.25} = \text{-1.61}$$

Finally we get the relevance value of the intent to the sentence:

$$Relevance\ value = 18 \text{ - } 1.61 = 16.39$$

4 Performance Measurements

We tested the performance of the system with 177 annotated emails. Each email in the corpus was classified into different folders representing intents. For example, an email containing an intention to confirm an appointment and request details about an appointment was filed in two folders named with the intent title. Thus, the 143 emails that contain intents in our corpus offer a total of 297 occurrences of the different intents.

We evaluated our system according to two criteria: its performance in email classification and the relevance of the answers it proposes. We made the same evaluation with other systems based on machine learning algorithms.

4.1 Effectiveness in Intents Detection

To compare our approach to machine learning approaches, we trained machine learning models with the 1150 annotated emails used to build the ontology and we tested them on the 177 emails corpus. We applied a Cross Validation method on predictive models using Logistic Regression, Decision Tree, Random Forest and Naive Bayes algorithms with a Bag of Words approach scored with TF-IDF. The Cross Validation suggests a best model based on Random Forest algorithm for the intent detection task.

We calculated precision, recall and f-score of the two systems based on the first intent proposition and the three intent propositions.

Table 2 and 3 show respectively the results of our system and of the machine learning based model.

Table 2: Precision, recall and f-score obtained by the ontology-based system.

	Precision	Recall	F-score
According to the 1st intent	0.50	0.48	0.49
According to the three intents	0.71	0.67	0.69

Table 3: Precision, recall and f-score obtained by the machine learning based model.

	Precision	Recall	F-score
According to the 1st intent	0.25	0.17	0.20
According to the three intents	0.25	0.23	0.24

For each evaluation step, our system presents far better results than the machine learning based model. In addition, we notice that our system captures more easily distinctions between the different intents (despite a decrease in the f-score caused by some confusion between intentions that share the same FEs and LUs) while machine learning has difficulties making multiclass distribution.

Table 4 shows the percentage of instances correctly identified by both systems for each intent. We compute these ratios according to the three proposed intents by each system since it is more likely to find the right intention on three proposals.

Our ontology-based system obtains the best results for almost all categories. The low results obtained by the machine learning system for all the categories can be explained by the size of the training corpus. Such a system requires many annotated examples to be able to learn and classify intents on its own. This experience shows instead that our approach is viable even with a small set of data.

The machine learning system obtained the best score on detecting the *precisions about an appointment* intent. It can be explained by the fact that many emails of the corpus contain this intent. A closer look at the results shows that the system tends to classify almost all emails with this intent. It is therefore logical that the proposed answer is often correct for this case.

Table 4: Effectiveness of intent detection for each intent type.

Intents	**Number of instances to find**	**Ontology based results**	**Ontology based ratio**	**ML based results**	**ML based ratio**
Request an appointment cancellation	2	2	**100%**	0	0%
Request an appointment confirmation	22	16	**73%**	13	59%
Request to schedule an appointment	8	3	**38%**	0	0%
Request an appointment change	3	0	0%	0	0%
Request details about an appointment	10	9	**90%**	1	10%
Request availabilities for an appointment	25	20	**80%**	6	24%
Request a participation to an appointment	7	3	**43%**	0	0%
Propose an appointment	39	30	**77%**	29	74%
Propose an appointment cancellation	2	1	**50%**	0	0%
Propose an appointment change	11	7	**64%**	0	0%
Availability confirmation	4	2	**50%**	2	**50%**
Appointment confirmation	8	4	**50%**	2	25%
Appointment cancellation notification	7	4	**57%**	0	0%
Unavailability notification	36	23	**64%**	6	17%
Availability notification	3	2	**67%**	1	33%
Appointment reminder notification	25	16	**64%**	3	12%
Precisions about an appointment	47	35	75%	42	**89%**
Appointment change notification	4	1	**25%**	0	0%

Anyway, our system seems able to correctly classify most of the intents. We notice that it does not identify the *request an appointment change* intent that is probably too close to the *propose an appointment*. In many cases, our system suggests a *propose an appointment* intent instead of a *request an appointment change* intent. Moreover, this case is not really a problem as long as given that the two intentions share similar answer templates. We do not combine these two intents into one because the *request* concept and the *proposal* concept are semantically different.

IC 2019

4.2 Evaluation of the Relevance of the Proposed Answers

We compared the relevance of the answers proposed by our system and by Google's Smart-Reply (Kannan *et al.*, 2016) – that uses Deep Learning and LSTM neural networks to propose 3 answers – for 15 emails. The relevance of the answers was evaluated by 26 Linagora employees through a questionnaire.

For each email, we asked annotators to choose between two blocks of answers – the Google's one and ours – the one that they felt contained the most relevant answer. When none of the proposed answers were relevant to the email submitted, annotators could select an option "no relevant answer". However, we did not allow annotators to select proposals from both systems at the same time even though the proposed answers were all relevant. We wanted to get as close as possible to a real situation where the user of an automatic answering system must necessarily choose one answer between several choices. We did not indicate which answers were proposed by the Google's system or by ours to obtain the most objective results possible.

Table 5 details the obtained results from 26 annotators by explaining the intents contained in each email submitted for annotation.

Table 5: Distribution of the chosen answers for each email.

Email	Contained intent	Ontology Based System	Google's Smart Reply	No relevant answer
1	Propose an appointment Request availabilities for an appointment	**15**	5	6
2	Request a participation to an appointment	5	**19**	2
3	Appointment confirmation	7	**18**	1
4	Request an appointment confirmation	**24**	1	1
5	Propose an appointment Request availabilities for an appointment	**16**	8	2
6	Propose an appointment cancellation	9	**17**	0
7	Precisions about an appointment	4	**20**	2
8	Appointment confirmation Request an appointment confirmation	5	**20**	1
9	Propose an appointment change	**13**	10	3
10	Propose an appointment cancellation	**20**	0	6
11	Request details about an appointment	**21**	5	0
12	Unavailability notification	**12**	**12**	2
13	Propose an appointment Availability notification Request availabilities for an appointment	6	**17**	3
14	Propose an appointment Request an appointment confirmation	**22**	4	0
15	Appointment cancellation notification	**23**	2	1
	Total:	202	158	30

The table shows that the answers obtained by our system are preferred for 8 emails out of 15 with an agreement of more than 75% for 5 emails against 7 emails out of 15 for the Google's system – with an agreement of more than 75% for only 2 emails. On the 12th mail, the same number of annotators chose the answers of the two systems. The total number of annotators' choices for responses from our system (202) is higher than Google's responses (158). We can also notice that our answers are still chosen by more than two annotators in all the cases for which the answers proposed by Smart-Reply are mainly selected by the annotators. In all cases, the answers proposed by our system were chosen by at least one

annotator which proves that these answers are grammatically correct. This experience shows that our approach allows us to provide answers that are relevant and acceptable to users. The results show that our answer proposals can fully compete with those proposed by the Google's system. A closer look at the annotators' choices also shows that the user prefers the most complete answers possible – that are proposed by our system. These are particularly chosen when the answer requires a mention of specific elements such as availabilities or precisions about an appointment. Figure 4 illustrates an example of sentence mainly chosen by annotators for an email containing an *appointment confirmation request* intent.

Received email:

"Re bonjour, je vous ai envoyé une invitation d'entretien pour mercredi 23 de 8h30 à 9h. Merci de bien vouloir la confirmer. "

(Good morning again, I sent you a meeting invitation for Wednesday the 23rd from 8:30 a.m. to 9:00 a.m. Thanks for confirming.)

Ontology-based system answers (24 choices):

1. "D'accord pour faire le rendez-vous mercredi 23 de 8h30 à 9h." (Okay to make the appointment Wednesday 23rd from 8:30 a.m. to 9:00 a.m.)
2. "Je ne serai malheureusement pas disponible pour ce créneau. Pouvons-nous trouver un autre créneau pour le rendez-vous ?" (Unfortunately, I will not be available for this time slot. Can we find another time slot for the meeting?)
3. "Je ne pourrai pas me libérer pour le rendez-vous mercredi 23 de 8h30 à 9h." (I won't be able to make an appointment Wednesday 23rd from 8:30 a.m. to 9:00 a.m.)

Google's Smart-Reply answers (1 choice):

1. "Bien reçu." (Alright)
2. "D'accord." (Okay)
3. "Bien reçu, merci." (Alright, thanks)

No relevant answer (1 choice)

Figure 4: Sample annotation for an email.

5 Advantages, Limitations and Further Research

Automatic email answering is a large studied problem. We propose a new way to resolve this problem by using an ontology to detect intent on emails and answer to them. Unlike machine learning solutions like Google's Smart-Reply our approach does not need large data set to run. Using an ontology is also an effective way to model complex semantic facts from text like meeting intents. Our system proposes answers more precise – thanks to the detection of complex intents – and complete with details on time, date or place of the meeting – thanks to named entities detection. Our approach is generic and new rules can easily be added to the ontology to detect new intents and answer to them. In our opinion, such a system is more controllable and understandable than a machine learning solution.

However, our approach has some limitations since the addition of intent concepts can be costly in time and thought. If the rules described by an intent concept are too precise, the system can easily merge several intents and provide noise. On the contrary, the system risks

missing many occurrences and thus generating silence if the rules associate too few lexical units and frame elements.

In addition, the effectiveness of this kind of symbolic rules depends on several factors:

- the domain: for instance, the meeting intents can be detected with the lexical units "table ronde" (a round-table discussion) or "colloque" (a symposium) in an academic setting, but rather with the lexical units "réunion" (a meeting) or "point" (a point) in a professional setting.
- the spelling: as the OpenPaaS mailbox does not integrate a spell checker, some lexical units may not be detected if misspelled. Moreover, it is difficult to manage all variations of declensions without an effective lemmatizer for French, which has many irregular verbs.
- synonym management: the rules must account for as many synonyms as possible to be able to recognize them all. A non-identified synonym cannot be recognized by the system.

We plan to improve our work by enriching the ontology with other intents (recruitment intents, reminder intents, document sharing intents, etc.). We'll improve the generation of email answers by detecting situations that require a formal form of address in emails – for example, when the user must address a hierarchical superior. We also plan to link the answering mechanism to the user's electronic agenda to automatically check availabilities dates and generate answers depending on the user's constraints.

Also, it seems important to us to find a way to automatically manage lexical units synonymy – by using synonym dictionary like WordNet for instance – and to integrate a French lemmatizer to improve the detection of lexical units and frame elements in emails.

Finally, it is possible to adapt the tool to detect intents in the other langages available in CoreNLP (English, German, French, Chinese and Spanish). At the ontology level, the user will have to add the POS properties of the new language and enrich the ontology with LUs and EFs in this language. Currently, the Java code support French and English. To process a new language, we need to integrate in the Intent Detector module the CoreNLP pipeline associated with this new language. The Answering Generator module requires a lemmatizer adapted to the language. It will be necessary to modify it according to the chosen langage.

6 Conclusions

In this paper, we presented an ontology-based approach for automatic email answering. The ontology was manually built from a corpus of French emails. It models meeting intent concepts using the FrameNet principles and different answer templates for each intent. The automatic email answering system contains four main modules to process the incoming email, spot intents with pattern matching, compute a score according to rules described in the ontology and propose answers.

We evaluated the effectiveness of the intents detection by calculating precision, recall and f-score measurements and compared the results with those obtained by machine-learning approaches. We also evaluated the relevance of the proposed answers by them with those proposed by the Google's Smart-Reply system.

In both cases, we get encouraging results with our system despite a limited annotated corpus. These results show that our approach is a suitable alternative to machine learning techniques (especially DNN) which are very demanding in annotated data.

We plan to enrich the ontology with other kinds of intents. We will improve the answering generation mechanisms by detecting the gender of the recipient and the T-V distinction – a linguistic formality in the French language which refers to "tu" and "vous" usages. A synonym dictionary and a lemmatizer will also be integrated to improve the matching of lexical units and frame elements in emails.

References

AUSTIN J. L. (1962). *How to do things with words: The William James Lectures delivered at Harvard University in 1955*. Oxford: Urmson.

CARVALHO V. R. (2008). *Modeling Intention in Email*. PhD thesis, Carnegie Mellon University, Language Technologies Institute, School of Computer Science.

FILLMORE C. (1976). Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, volume 280, p. 20–32.

KANNAN A., KURACH K., RAVI S., KAUFMAN T., MIKLOS B., CORRADO G., TOMKINS A., LUKÁCS L., GANEA M., YOUNG P. & RAMAVAJJALA V. (2016). Smart reply: Automated response suggestion for email. In *Conference on Knowledge Discovery and Data Mining (KDD)*. San Francisco, USA.

KATAKIS I., TSOUMAKAS G. & VLAHAVAS I. (2006). Email mining: Emerging techniques for email management. In A. VAKALI & G. PALLIS, Eds., *Web Data Management Practices: Emerging techniques and Technologiques*, p. 219–240. Idea Group Publishing, USA.

KOSSEIM L., BEAUREGARD S. & LAPALME G. (2001). Using information extraction and natural language generation to answer e-mail. *Data Knowledge Engineering, vol.38*, p. 85–100.

MALIK R., SUBRAMANIAM L. V. & KAUSHIK S. (2007). Automatically selecting answer templates to respond to customer emails. In *Proc. 20th International Joint Conference on Artificial Intelligence (IJCAI)*, volume 7, p. 1659–1664.

MANNING C. D., SURDEANU M., BAUER J., FINKEL J., BETHARD S. J. & MCCLOSKY D. (2014). The standford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations.*, p. 55–60.

RUPPENHOFER J., ELLSWORTH M., PETRUCK M., JOHNSON C., BAKER C. & SCHEFFCZYK J. (2016). Framenet ii: Extended theory and practice. Available online: https://framenet.icsi.berkeley.edu/fndrupal/the_book.

SEARLE J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*, volume 626. Cambridge University Press.

SNEIDERS E. (2010). Automated email answering by text pattern matching. In H. LOFTSSON, E. RÖGNVALDSSON & S. HELGADÓTTIR, Eds., *Proc. 7th International Conference on Natural Language Processing (IceTAL), August 16-18, Reykjavik, Iceland, LNAI 6233*, p. 381–392. Springer, Berlin, Heidelberg.

Pourquoi dois-je croire ta prédiction ? Comment expliquer les résultats d'une classification automatique de sentiments à partir de textes

Waleed Ragheb[1,2], Jérôme Azé[1,2], Sandra Bringay[1,3], Maximilien Servajean[1,3]

[1] LIRMM UMR 5506, CNRS, University of Montpellier

[2] IUT DE BÉZIERS, University of Montpellier, Béziers

[3] AMIS, Paul Valéry University - Montpellier 3

Résumé : Dans le cadre d'un problème classique de classification de sentiments, nous proposons un modèle qui utilise 1) l'apprentissage par transfert plutôt que les méthodes classiques de word embedding et 2) des mécanismes d'attention permettant de se concentrer sur les parties du texte importantes pour la tâche de classification étudiée. Notre modèle a été évalué sur plusieurs jeux de données et montre des résultats très compétitifs. Or, si ces méthodes d'apprentissage en profondeur s'avèrent très efficaces, elles sont souvent considérées comme des boites noires, difficiles à interpréter. Dans cet article, nous évaluons l'impact des mécanismes d'attention traduits sous la forme de nuages de mots-clés pour aider les utilisateurs à interpréter les résultats de la classification. L'expérimentation d'une telle visualisation sur plus de 85 participants a permis de montrer son intérêt en terme d'interprétabilité.

Mots-clés : classification de sentiments, mécanisme d'attention et interprétabilité.

1 Introduction

L'analyse de sentiments à partir de textes est un domaine de recherche très actif dans la communauté de l'apprentissage automatique (Mäntylä *et al.*, 2018) avec de nombreuses applications comme le e-commerce, la gestion de la réputation, le support client, la politique, etc. De nombreux outils sont désormais disponibles pour détecter des sentiments subjectifs, tels que la polarité (positif ou négatif) et les émotions (peur, joie, etc.). Si ces outils sont de plus en plus efficaces, ils sont généralement critiqués pour deux raisons.

Tout d'abord, la plupart des approches existantes ne sont efficaces que pour des domaines d'application spécifiques ou des types de textes particuliers tels que l'analyse de sentiments des sites Web liés à la finance (Luo *et al.*, 2018) ou des avis clients relatifs à des ordinateurs portables ou des restaurants (Li *et al.*, 2018). Dans ce contexte, l'apprentissage par transfert et l'adaptation au domaine sont largement utilisés, en particulier combinés à des réseaux de neurones profonds, pour aider à réutiliser les modèles développés pour une tâche source vers une autre tâche cible. L'apprentissage par transfert fonctionne particulièrement bien quand les caractéristiques apprises pour la tâche source sont générales et peuvent être réutilisées pour les tâches cibles. Ce type d'approche a fait ses preuves dans le domaine de la vision par ordinateur où l'extraction des caractéristiques se fait à partir de modèles pré-entraînés de type AlexNet, ResNet, MS-COCO, etc. (Voulodimos *et al.*, 2018). Dans les modèles de traitement du langage naturel, cette approche n'a connu de réel succès que très récemment grâce au modèle de langage universel (ULMFiT) proposé par (Howard & Ruder, 2018) qui servira de baseline à ces travaux.

Par ailleurs, en traitement automatique de la langue, la plupart des modèles de transduction compétitifs ont une structure de type codeur-décodeur (Vaswani *et al.*, 2017). Une limite de ces architectures est qu'elle code la séquence d'entrée dans une représentation interne de longueur fixe. Cela entraîne une dégradation des résultats lorsque la longueur de la séquence

augmente. Dans ce contexte, les mécanismes d'attention (Young *et al.*, 2018) ont récemment été utilisés pour résoudre ce problème. Inspirés des mécanismes d'attention visuelle que l'on retrouve chez l'homme, ils focalisent l'analyse sur certaines régions d'une image avec une "haute résolution" tout en percevant le reste de l'image en "basse résolution". L'attention guide le réseau pour qu'il sache où accorder son attention sur la séquence d'entrée. Les premières applications des mécanismes d'attention ont été naturellement réalisées dans le domaine de la vision par ordinateur (Anderson *et al.*, 2017) et plus récemment sur les textes pour des application de traduction automatique (Bahdanau *et al.*, 2014) et d'analyse des sentiments (Ma *et al.*, 2018).

Dans ce travail, nous évaluerons la combinaison des mécanismes d'attention à une architecture de type ULMFiT sur des jeux de données réelles de la littérature.

Par ailleurs, nous nous poserons la question de l'interprétabilité de notre approche. Les modèles d'apprentissage sont souvent décriés car perçus comme des boites noires. Or, pour un utilisateur qui va baser des décisions et des actions à partir de prédictions, il peut être fondamental d'interpréter les raisons sous-jacentes aux prédictions pour ainsi faire confiance à ces prédictions. En effet, il y a des cas où l'erreur a peu d'importance, par exemple quand on recommande un film, mais il y a des cas où l'erreur porte à conséquence, par exemple pour le diagnostic médical ou la détection d'obstacles pour une voiture autonome. Selon l'importance de l'erreur pour l'activité de l'utilisateur, celui-ci aura besoin de comprendre comment le modèle fonctionne en général et comment il est arrivé à proposer une prédiction particulière.

Dans ce travail, nous nous intéressons plus particulièrement à la sortie des mécanismes d'attention qui est la visualisation des parties du texte ayant impacté la prédiction du sentiment. Cette visualisation donne des justifications complémentaires aux étiquettes prédites associées aux textes. À notre connaissance, aucune étude n'a montré l'apport de la visualisation des mécanismes d'attention sur l'interprétation des utilisateurs pour l'analyse de sentiments à partir de textes.

Finalement, l'objectif de cet article est double : 1) montrer l'impact en terme d'exactitude des mécanismes d'attention pour l'analyse des sentiments lorsqu'ils sont ajoutés à l'architecture de référence ULMFiT qui intègre un apprentissage par transfert ; 2) évaluer dans quelle mesure une visualisation basée sur le mécanisme d'attention facilite la prise de décision en apportant un élément explicatif supplémentaire à l'étiquette proposée.

Le reste de cet article est organisé comme suit. Dans la section 2, nous décrivons les travaux connexes de la littérature puis le modèle proposé dans la section 3. Dans la section 4, nous présentons les expérimentations sur les performances du modèle, les jeux de données et les résultats obtenus. Dans la section 5, nous détaillons les expérimentations réalisées sur l'interprétabilité du modèle. Nous concluons et donnons des perspectives dans la section 6.

2 État de l'art

Afin d'améliorer l'efficacité des algorithmes d'analyse de sentiments et leur explicabilité, nous proposons une nouvelle méthode basée sur l'apprentissage par transfert de modèles de langue ainsi que sur les mécanismes d'attention.

Les modèles de langue (LM) visent à prédire un mot à partir des mots le précédant. Ces modèles sont utilisés dans de nombreuses applications de traitement automatique de la langue naturelle car ils permettent de capturer des dépendances éloignées ainsi que la structure hiérarchique du texte. L'apprentissage des modèles de langue est non supervisé, car ne nécessitant pas de corpus de texte préalablement étiqueté. Or, ces modèles sont peu adaptés aux petits ensembles de données et peuvent donner un mauvais rappel pour certaines tâches de classification. Récemment, (Howard & Ruder, 2018) ont proposé la méthode ULMFiT, basée sur une représentation des mots peu profonde, qui combinée à un apprentissage par transfert s'est avérée être très efficace pour différentes tâches dont l'analyse de sentiments (Merity *et al.*, 2018). Cette méthode servira de baseline à nos travaux. Nous avons choisi d'utili-

ser ULMFiT pour ses performances en *fine tuning* et sa taille raisonnable en comparaison à d'autres approches très récentes comme BERT (Devlin *et al.*, 2018) et ELMO (Peters *et al.*, 2018). De plus, pour des tâches de classification de sentiments, ULMFiT est utilisé comme l'état de l'art sur la plupart des jeux de données de la littérature, dont ceux sur lesquels nous avons travaillé.

Dans le domaine du traitement automatique de la langue naturelle, les mécanismes d'attention ont été récemment étudiés. L'auto-attention met en relation différentes positions d'une séquence afin d'en calculer une représentation (Lin *et al.*, 2017). Chaque partie de la séquence d'entrée est associée à un score de probabilité. L'attention a été appliquée avec succès pour de nombreuses tâches, notamment la compréhension de la lecture, le résumé, la traduction automatique (Su *et al.*, 2018).

Outre le fait que les mécanismes d'attention améliorent l'efficacité de la classification, ce qui nous intéresse dans cet article est le fait que le résultat de la couche d'attention peut être utilisé en entrée d'une visualisation visant l'explication d'une prédiction (Wang *et al.*, 2018) (Lin *et al.*, 2017).

Une explication est une réponse à une question de type « Pourquoi ? ». Dans notre cas, pourquoi le système prédit une polarité pour un texte ? Il a été démontré par (Herlocker *et al.*, 2000) qu'associer des explications aux prédictions améliore l'acceptation de ces prédictions dans le cas de la recommandation de films. En effet, les explications influencent le ressenti de celui à qui l'on fournit les explications et donc les actions qu'il peut être amené à réaliser. Pour expliquer une prédiction, la plupart des méthodes utilisent des artefacts visuels qui fournissent une compréhension qualitative des liens existants entre les instances (des mots, des parties d'images...) et les prédictions.

Il existe différentes méthodes d'interprétabilité. On distingue les méthodes d'apprentissage intrinsèquement interprétables lorsque la sélection et l'entraînement du modèle d'apprentissage sont par eux-même interprétables (e.g. les arbres de décision) et les méthodes d'interprétabilité post-hoc qui donnent des explications *a posteriori* et qui s'appliquent sur des méthodes d'apprentissage de type boite noire après la sélection et l'entraînement du modèle, pour expliquer les prédictions réalisées. La plupart des modèles d'apprentissage de l'état de l'art dont les modèles de type réseau de neurones étudiés dans cet article, n'étant pas interprétables, nous nous sommes focalisés sur cette deuxième catégorie de méthodes d'interprétabilité.

À notre connaissance, il n'y a pas eu d'étude qualitative de l'impact de la visualisation de l'attention sur l'interprétation que les utilisateurs peuvent faire de cette nouvelle information dans le cas de la classification de données textuelles. Cette étude nous parait importante dans le contexte du besoin de confiance manifesté par les utilisateurs des méthodes d'apprentissage notamment dans des domaines comme la santé. Ces utilisateurs ne sont pas uniquement en attente de résultats performants mais ils recherchent des explications qui, associées aux prédictions, vont améliorer la prise de décision (Ribeiro *et al.*, 2016)(Vellido *et al.*, 2012).

3 Architecture proposée

Notre architecture est composée de quatre composants décrits dans les sections suivantes.

3.1 Auto-Attention basée sur un encodeur AWD-LSTM

Un LSTM traditionnel possède un portail d'entrée i_t, d'oubli f_t, de sortie o_t et une cellule mémoire c_t. Ce sont des vecteurs de $\mathbb{R}^d$ qui correspondent à la représentation vectorielle

d'une dimension d. Les équations de transition de LSTM sont les suivantes :

$$\begin{aligned} f_t &= \sigma(W_f x_t + U_f h_{t-1}) \\ i_t &= \sigma(W_i x_t + U_i h_{t-1}) \\ o_t &= \sigma(W_o x_t + U_o h_{t-1}) \\ \tilde{c}_t &= tanh(W_c x_t + U_c h_{t-1}) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\ h_t &= o_t \odot tanh(c_t) \end{aligned} \tag{1}$$

où x_t est l'entrée au pas de temps actuel, σ est la fonction sigmoïde ; $\odot$ l'opération de multiplication par élément, $W_{\{i,f,o,c\}}$, $U_{\{i,f,o,c\}}$ sont des ensembles de poids appris.

Dans notre modèle, nous utilisons le vecteur d'état caché à chaque pas de temps comme représentation du mot correspondant dans une phrase. Afin d'éviter le sur-apprentissage lors de l'entraînement du LSTM, (Merity *et al.*, 2018) ont proposé AWD-LSTM. basée sur DropConnect (Wan *et al.*, 2013) pour pondérer les matrices $U_{\{i,f,o,c\}}$. Nous avons utilisé les trois mêmes couches liées de LSTM et avons également appliqué une auto-attention sur les vecteurs d'état cachés à chaque pas de temps. La séquence d'états cachés en entrée $H^{i-1} = \{h_1^{i-1}, h_2^{i-1}, \ldots, h_N^{i-1}\}$, où N est la longueur de la séquence, est passée aux états de la couche LSTM. Les états de sortie sont de la forme de $H^i = \{h_1^i, h_2^i, \ldots, h_N^i\}$. La couche d'attention prend la séquence en entrée encodée et calcule les scores d'attention $S^i = \{s_1^i, s_2^i \ldots, s_N^i\}$. La couche d'attention est une couche linéaire sans biais.

$$\begin{aligned} \alpha^i &= \{V^i . H^i\} \\ S^i &= exp(\alpha^i) / \sum_{j=1}^{N} exp(\alpha_j^i) \end{aligned} \tag{2}$$

Où V^i est le poids de la couche d'attention i^{th} du Att-LSTM.

3.2 Agrégation sur plusieurs niveaux de l'auto-attention

L'architecture proposée utilise un empilement de trois Att-LSTM superposés exactement, de la même manière que l'architecture AWD-LSTM classique. Nous n'utilisons pas d'architecture Bi-LSTM ici, car notre modèle correspond à l'ensemble des directions avant et arrière. La figure 1 montre un aperçu du modèle. À chaque couche, les scores d'attention sont obtenus selon un niveau spécifique de codage de la séquence, puis agrégés pour obtenir les scores d'attention globaux $\overline{S}$. La fonction d'agrégation est la moyenne logarithmique des trois niveaux de scores d'attention.

$$\overline{S} = log \sum_{i=1}^{3} S^i / 3 \tag{3}$$

Les scores globaux d'attention $\overline{S}$ sont utilisés pour calculer la séquence $O = \{o_1, o_2, \ldots, o_N\}$ où o_i est le produit du score d'attention et de la sortie de la couche Att-LSTM, tel que :

$$o_i = \overline{s}_i \otimes h_i^3 \tag{4}$$

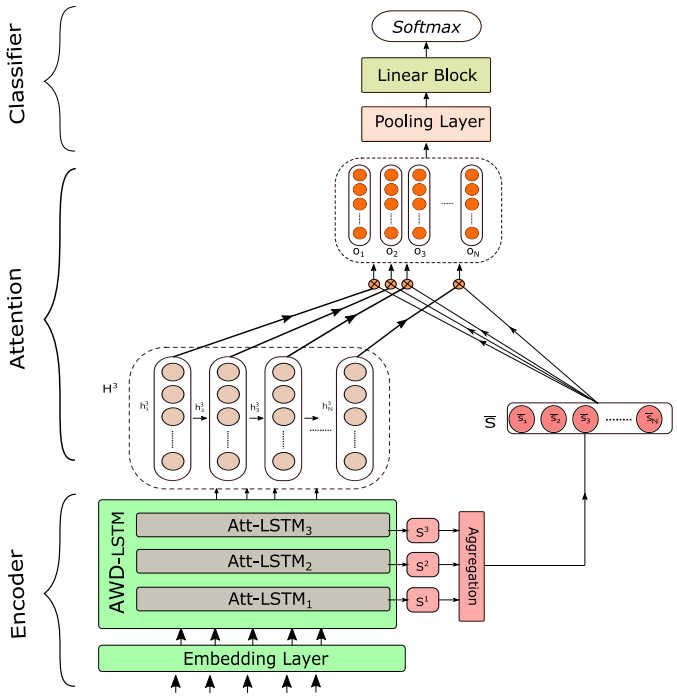


FIGURE *1 – Notre architecture*

3.3 Couches de classification

Après avoir agrégé les informations provenant de l'attention multi-niveaux et de la sortie de l'encodeur, nous convertissons les représentations résultantes de toutes les positions de O en un vecteur à longueur fixe avec une opération de pooling. Nous avons utilisé trois fonctions de pooling. Nous appliquons une attention X_{att} telle que :

$$X_{att} = \sum_{i=1}^{N} exp(\overline{s}_i) \otimes h_i^3 \tag{5}$$

Nous appliquons un pooling maximum X_{max} et moyen X_{avg} à O afin d'obtenir les représentations finales du texte saisi après encodage et application de l'attention, donné par :

$$X_{in} = [X_{att} \oplus X_{max} \oplus X_{avg}] \tag{6}$$

Puis, nous alimentons le bloc linéaire du classifieur. Ce bloc est constitué de deux couches entièrement connectées, de tailles différentes, suivies d'un indicateur *softmax* pour déterminer la classe de sentiment en sortie.

3.4 entraînement du modèle

L'entraînement se fait en trois étapes :

— Le modèle de langue est initialisé de manière aléatoire, puis entraîné en empilant un décodeur linéaire au-dessus de l'encodeur. Le modèle de langue est appris à partir d'un corpus du domaine général. Cela aide le modèle à apprendre des caractéristiques générales du langage.

— Le même modèle de langue après entraînement est utilisé pour l'initialisation puis ajusté à partir des données de la tâche cible, dans notre cas : différents jeux de données d'analyse de sentiments. Ici, nous limitons le vocabulaire du modèle de langue aux mots fréquents (répétés plus de deux fois).

— Nous conservons l'encodeur et remplaçons le décodeur par le classifieur et les paramètres de ces deux étapes sont réglés avec précision pour la tâche cible.

Lors de la première étape et pour l'apprentissage du modèle de langage, nous avons utilisé le jeu de données Wikitext-103 (Merity *et al.*, 2016). Avec plus de 28 000 articles Wikipédia et 103 millions de mots, le modèle détermine la structure principale et la hiérarchie du langage par modélisation séquence par séquence. Pour l'entraînement du modèle de classification, à partir du jeu étiqueté, nous optimisons tous les paramètres afin de minimiser autant que possible la fonction objectif (fonction de perte). Dans notre travail, nous prenons y_i la polarité du sentiment correcte et $\hat{y_i}$ qui désigne la polarité de sentiment prédite. Nous considérons l'entropie croisée comme la fonction de perte, dont la formule est la suivante :

$$loss = -\sum_{<T>} y_i \log(\hat{y_i}) + \lambda \|\theta\|^2 \tag{7}$$

Où λ est le facteur de régularisation, θ contient tous les paramètres du modèle et T est l'ensemble des exemples d'entraînement. L'entraînement de l'architecture est effectué à l'aide de taux d'apprentissage triangulaires inclinés (STLR) qui modifient le taux d'apprentissage pour chaque itération de manière triangulaire. Nous avons utilisé un seul cycle comme recommandé par (Howard & Ruder, 2018). Le modèle a été entraîné en utilisant un taux d'apprentissage différent pour chaque groupe de couches.

Nous entraînons le modèle sur les modèles de langue avant et arrière pour les ensembles de données du domaine général et spécifiques à la tâche. Les deux modèles de langue avant et arrière sont utilisés pour construire deux versions de la même architecture. La décision finale est l'ensemble des deux. Nous avons utilisé Pytorch [1] pour construire l'ensemble du modèle et les bibliothèques Fastai [2] pour l'apprentissage et les modèles de langue. Pour le prétraitement du texte, celui-ci a tout d'abord été normalisé et tokenisé. Des tokens spéciaux ont été ajoutés pour les mots en majuscules et les mots répétés. Nous conservons les symboles de ponctuation et de sentiments dans le texte. Nous avons utilisé Spacy [3] et FastText [4] pour ces prétraitements. Les modèles sont appris et testés sur 4 GPU Nvidia GEFORCE GTX 1080 ti.

4 Expérimentations sur les performances du modèles

Dans cette section, nous discutons plus en détail de l'efficacité de la méthode proposée.

4.1 Jeux de données

Nous appliquons le modèle à différents jeux de données de classification de sentiments. Le tableau 1 présente des statistiques sur ces jeux de données. Le jeu de données IMDB est un ensemble de données pour la classification des sentiments binaires de critiques de films (Maas *et al.*, 2011). Nous avons également utilisé les versions binaires et complètes des jeux de données d'avis d'utilisateurs de Yelp et d'Amazon (Zhang *et al.*, 2015). Pour les jeux de données binaires (IMDB, Yelp-bi et Amazon-bi), les classes à prédire sont positif ou négatif. Pour les autres, il s'agit d'un nombre d'étoiles : de négatif (1 étoile) à positif (5 étoiles). Ces jeux de données sont tous équilibrés.

4.2 Baselines et Résultats

Nous comparons notre modèle à plusieurs baselines compétitives de l'état de l'art qui utilisent le mécanisme d'attention pour la classification de sentiments :

1. https://pytorch.org/
2. http://www.fast.ai/
3. https://spacy.io/
4. https://fasttext.cc/

TABLE *1 – Jeux de données de sentiments et nombre d'exemples d'apprentissage et de test*

Dataset	#Exemples d'apprentissage	#Exemples de test	#classes
IMDB	25K	25K	2
Yelp-bi	560K	38K	2
Yelp-Full	650K	50K	5
Amazon-bi	3.6M	400K	2
Amazon-Full	3M	650K	5

TABLE *2 – Taux d'erreur (%) de notre modèle et des baselines*

Models	IMDB	Yelp-bi	Yelp-Full	Amazon-bi	Amazon-Full
HN-ATT	-	-	-	-	36.40
DCCNN-ATT	-	2.64	30.58	**3.32**	34.81
SANet	-	4.77	36.03	4.52	38.67
SA-Embedding	-	5.10	36.60	-	40.20
CSC	-	6.90	35.97	4.90	39.89
CRAN	7.90	-	-	-	-
IRAM	8.80	-	-	-	-
ULMFiT	4.60	**2.16**	29.98	-	-
Ours	**4.51**	2.25	**29.76**	3.43	**34.78**

— HN-ATT (Yang *et al.*, 2016) Le modèle reflète la structure hiérarchique des documents à travers deux niveaux d'attention dans les mots et les phrases.
— DCCNN-ATT (Wang *et al.*, 2018) Ce modèle est un réseau de neurones convolutifs avec des connexions denses et des fonctionnalités multi-échelles.
— SANet (Letarte *et al.*, 2018) Ce modèle utilise l'attention pour modéliser les interactions entre toutes les paires de mots d'entrée.
— SA-Embedding (Lin *et al.*, 2017) Ce modèle est basé sur l'extraction d'une représentation interprétable de la phrase donnée en entrée via un mécanisme d'attention.
— CSC (Mokhtari *et al.*, 2018) Ce modèle utilise un réseau de neurones hiérarchique basé sur l'attention qui intègre les préférences de l'utilisateur et les caractéristiques du produit dans les tâches de classification de sentiments.
— CRAN (Du *et al.*, 2017) Le modèle associe à la fois les attentions basées sur la convolution et les attentions basées sur la récurrence.
— IRAM (Tutek & Šnajder, 2018) Il s'agit d'un modèle d'attention, qui construit de manière récursive des représentations d'entrée des données par la réutilisation des résultats qui ont été précédemment calculés.
— ULMFiT (Howard & Ruder, 2018) Il s'agit de la méthode de référence actuelle .

Le tableau 2 montre l'erreur obtenue lors du test du modèle proposé et de toutes les baselines sur les jeux de tests. Nous présentons les résultats tels que rapportés dans la publication d'origine. Notre modèle surpasse tous les modèles basés sur l'attention avec une marge significative et reste compétitif par rapport à ULMFiT. Notre modèle proposé obtient de meilleurs résultats qu'ULMFiT pour le jeu IDBM et la version complète de Yelp.

Dans la suite, nous allons nous demander comment améliorer l'interprétabilité de ce modèle et montrer comment les mécanismes d'attention peuvent être utilisés pour expliquer les prédictions.

5 Expérimentations sur l'interprétabilité du modèle

Dans cette section, nous discutons plus en détail de l'impact de l'application de l'attention sur l'interprétabilité.

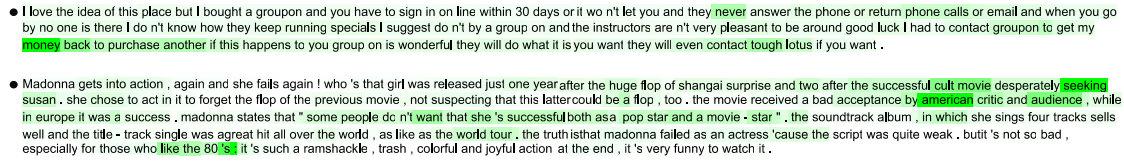


FIGURE *2 – Exemples de visualisation de l'attention sur une critique de restaurant et de film*

5.1 Visualisation de l'attention

L'utilisateur doit avoir confiance dans un modèle et pour cela, il voudra vérifier que celui-ci fonctionne bien sur des données réelles selon une métrique d'intérêt, comme le taux d'erreur utilisé dans la section 4. Ce type d'évaluation peut s'avérer inefficace notamment quand les données évoluent au cours du temps car elle revient à évaluer entre autre l'écart entre les données réelles et celles d'entraînement. Un humain peut alors élaborer différentes stratégies pour sélectionner un modèle parmi plusieurs modèles et notamment préférer un modèle explicable mais moins performant, qui lui laisse la responsabilité du choix final.

Dans ce contexte, l'un des résultats les plus intéressants du mécanisme d'attention est sa capacité à traiter toutes les séquences d'entrée avec différents poids d'attention. Le modèle accorde une plus grande attention aux éléments qui influencent la décision du réseau. La figure 2 montre des exemples d'avis positifs et négatifs correctement classés pour les avis de restaurants (Yelp-bi) et de critiques de film (IDBM). Les scores d'attention sont utilisés pour colorer le texte. Cette information peut être alors présentée à l'utilisateur avec la prédiction comme explication.

Dans notre contexte, une explication est un ensemble de mots-clés associés à un score d'attention. Dans la suite, nous allons chercher à évaluer la qualité des explications fournies via la visualisation de l'attention selon une approche centrée sur le système (le niveau d'interprétabilité est basé sur l'analyse des sorties du système) puis selon une approche centrée sur l'humain (le niveau d'interprétabilité est basé sur une tâche expérimentale d'inférence de la polarité).

5.2 Évaluation centrée système

Nous avons tout d'abord mis en place une évaluation centrée système, sans intervention humaine. Pour cela, nous avons mesuré l'intersection entre le nombre de mots appartenant à des lexiques ayant fait leur preuve pour les tâches de classification de sentiments et les mots repérés par les mécanismes d'attention. Nous avons utilisé pour cela EmoLex proposé par (Mohammad & Turney, 2013) (Mohammad, 2018) qui comporte plus de 10 000 entrées.

Le tableau 3 montre les résultats de la mise en correspondance des principaux mots repérés avec l'attention et retrouvés dans EmoLex pour les deux jeux de tests. Par exemple, nous pouvons dire que 88,11% des exemples du jeu de données IMDB contiennent des sentiments dans les 5% supérieurs des mots identifiés par le processus d'attention. Cela reflète la précision du mécanisme d'attention qui se concentre sur ces mots et expressions.

TABLE *3 – Sentiments et émotions dans les top 5, 10 ou 20% des scores d'attention calculés*

Dataset	Top 5%	Top 10%	Top 20%
IMDB	88.11%	97.30%	99.65%
Yelp-bi	51.86%	78.81%	94.37%
Yelp-Full	64.11%	84.65%	95.47%
Amazon-bi	55.71%	80.18%	94.65%
Amazon-Full	57.67%	81.02%	94.84%

FIGURE *3 – Exemple de nuage de mots-clés généré à partir des mécanismes d'attention*

5.3 Évaluation centrée humain

Même si les modèles d'attention se concentrent sur certains mots porteurs d'émotions, ils se concentrent aussi sur d'autres parties du texte, ayant un impact important sur la décision finale du réseau. Sur la base d'exemples tels que ceux de la figure 2, nous pourrions facilement trouver des parties du texte surlignées qui ne sont pas liées aux lexiques de sentiments.

Afin de déterminer dans quelle mesure une explication basée sur le simple fait de prêter attention aux mots importants, malgré une sélection imparfaite par l'algorithme, pourrait induire une décision de polarité meilleure qu'une explication basée sur les termes d'un lexique, nous avons mis en place une évaluation centrée humain impliquant des utilisateurs non experts. Pour cela, nous leur avons présenté les deux types d'explications générées sous la forme de nuages de mots comme présenté sur la figure 3.

Les nuages de mots sont générées de trois manières différentes : en utilisant 1) uniquement le lexique, 2) uniquement l'attention, ou 3) les deux. Dans le cas des nuages de mots générés à partir du lexique, la taille des mots est déterminée par les fréquences des mots et leur ordre dans le texte. Dans le cas des mécanismes d'attention, la taille et la position des mots sont déterminées par le score d'attention donné dans la section 3.1 . Pour les nuages de mots mixtes, nous combinons les deux informations.Les couleurs des mots sont choisies pour augmenter le contraste entre les mots de tailles différentes pour une meilleure visualisation. Ces couleurs ne représentent aucune information supplémentaire.

Chaque participant a répondu à 30 questions, 10 questions pour chaque modalité. La question posée a consisté à deviner la polarité des critiques (positives ou négatives) en se basant sur les nuages de mots. Nous informons le participant si le nuage de mots provient d'un commentaire de film, de restaurant ou d'un autre produit. La sélection des questions est randomisée pour chaque type de question et est différente pour chaque participant. Le nombre total de participants est de 85. Il s'agit d'étudiants inscrits dans des cursus en informatique. Pour chaque type de question, nous calculons l'index moyen de facilité F_i comme suit :

$$F_i = 100 * \frac{\bar{x_i} - x^i_{min}}{x^i_{max} - x^i_{min}} \tag{8}$$

où $\bar{x_i}$ est le score moyen par question de type i, x^i_{min} et x^i_{max} sont les scores minimum et maximum pour une question de type i.

Le score par question donne une indication sur la facilité avec laquelle la question est posée. Nous calculons également l'écart-type de tous les scores pour chaque question. De plus, nous calculons l'indice de discrimination D_i comme :

$$D_i = 100 * corr(x_i, T) \tag{9}$$

qui correspond au pourcentage de corrélation entre les scores de chaque type de question et le score total de chaque participant. Le tableau 4 présente les résultats du sondage sur les nuages de mots pour les trois types de questions.

TABLE *4 – Résultats de l'enquête sur les nuages de mots*

Types de Question	Index de facilité	Déviation Standard	Index de discrimination
Attention	**72.12%**	**43.90%**	**7.70%**
Lexicon	58.90%	49.86%	-16.10%
Mixte	68.27%	49.19%	6.70%

Ces résultats restent préliminaires. Ils montrent que deviner le sentiment sans le texte complet est une tâche difficile. Cependant, les questions basées sur les nuages de mots construits par les mécanismes d'attention sont plus faciles que les autres avec un indice de facilité moyen de 72.12% et un écart-type de 43.90%. Par ailleurs, nous notons une corrélation positive entre l'accord et l'indice de discrimination. Nous pouvons en conclure que la tâche devient plus difficile en utilisant uniquement le lexique. Toutefois, les nuages construits à partir des scores d'attention captent des arguments qui ne sont pas centraux à la tâche et ne mettent pas en avant les expressions de négation et adversatives qu'il serait intéressant de repérer et de mettre en relief dans cette visualisation. D'autres formats de visualisation pourraient également être explorés.

5.4 Discussions

La méthode d'interprétabilité proposée, basée sur les mécanisme d'attention, permet de visualiser un résumé des statistiques (taille et ordre) des attributs (les mots) et de leur impact sur les prédictions du modèle (intensité de la couleur). Cette méthode d'explication est spécifique à l'algorithme d'apprentissage utilisé qui fournit l'information de l'attention et ne rentre donc pas dans la catégorie des modèles d'interprétation *a posteriori*, qui peuvent être utilisés avec n'importe quel autre algorithme d'apprentissage. Un avantage de notre méthode est qu'elle est expressive car la représentation visuelle retenue permet de structurer les explications. Comme les auteurs de la méthode LIME (Hu *et al.*, 2018), nous avons pu remarquer que les utilisateurs apprécient les explications qu'ils sélectionnent par eux-mêmes. En effet, ils ne s'attendent pas à une liste finie d'explications de la prédiction mais plutôt au choix d'une ou plusieurs explications dans une liste de choix, ce qui est possible en piochant des mots-clés dans le nuage. Toutefois, une limite est que l'interprétation proposée est locale et se limite à une prédiction unique alors que d'autres approches dites globales vont concerner tout le modèle d'apprentissage et notamment le choix des paramètres.

(Molnar, 2019) a listé les propriétés que l'on peut rencontrer pour évaluer la qualité des explications individuelles. Il considère :

— la précision (*dans quelle mesure une explication permet-elle de prédire des données non encore rencontrées ?*) : contrairement aux approches par lexique, les explications liées à l'attention resteront efficaces même avec un vocabulaire non standard,
— la fidélité (*dans quelle mesure l'explication se rapproche-t-elle de la prédiction du modèle vu comme une boîte noire ?*) : l'expérimentation a montré que l'humain est capable de retrouver la prédiction à partir des nuages,
— la cohérence (*en quoi une explication diffère-t-elle entre les modèles entraînés à la même tâche et produisant des prédictions similaires ?*) : le nuage de mots permet par exemple d'identifier les raisons pour lesquelles deux critiques de films vont être considérées comme positives (en soulignant par exemple des termes liés à la qualité de la réalisation ou au jeu des acteurs),
— la stabilité (*dans quelle mesure les explications sont-elles similaires pour des instances similaires* ?) : pour des critiques similaires, les mêmes mots sont repérés par l'attention,

— la compréhensibilité (*dans quelle mesure les humains comprennent-ils les explications ?* : même si la tâche a été considérée comme difficile, les humains ont réussi à prendre une décision à partir des nuages et ont trouvé la bonne prédiction avec un index de facilité de 72.12%,
— la certitude (*l'explication reflète-t-elle la certitude du modèle d'apprentissage automatique ?*) : le nuage de mots ne retourne pas d'information sur la certitude que le modèle a en sa prédiction mais cette information pourrait être ajoutée sous la forme d'une nouvelle variable visuelle,
— le degré d'importance (*dans quelle mesure l'explication reflète-t-elle l'importance de caractéristiques ou de parties de l'explication ?*) : cette information est représentée par la taille des mots dans le nuage,
— la nouveauté (*L'explication indique-t-elle si une instance à expliquer provient d'une région très éloignée de la distribution des données d'entraînement ?*) : la visualisation actuelle ne donne pas cette information.

Notre modèle d'explication locale des prédictions basé sur les nuages de mots-clés repérés par l'attention possède donc des propriétés intéressantes au sens de (Molnar, 2019).

6 Conclusions et perspectives

Nous pouvons conclure de cette étude que l'idée de l'apprentissage par transfert est très efficace pour une application de classification de sentiments en terme de taux d'erreur. Elle fonctionne mieux que les modèles classiques d'apprentissage basés sur les word embeddings. De plus, l'ajout d'un mécanisme d'auto-attention a un impact direct sur les performances de ces modèles. Le modèle proposé a été évalué sur cinq jeux de données de la littérature. Nos expériences montrent des résultats compétitifs par rapport aux modèles basés sur l'attention à la pointe de l'état de l'art dont ULMFit. Pour finir, même si la visualisation proposée peut facilement être améliorée et que les expérimentations restent préliminaires, nous avons démontré que la visualisation des scores d'attention présentés sous la forme de nuages de mots, impacte positivement les perceptions liées à l'interprétabilité des utilisateurs.

Dans les travaux futurs, une nouvelle expérimentation, cette fois comparative, devra être menée pour savoir si l'ajout d'une information comme la visualisation proposée permet effectivement à un utilisateur d'avoir confiance dans le modèle et si cela impacte sa prise de décision. Nous allons chercher à améliorer l'interprétation locale d'une prédication. Par exemple, il a été démontré que les humains apprécient les explications contrastives, qui permettent de comprendre pourquoi une prédiction a été faite à la place d'une autre. Pour un texte pour lequel nous cherchons à déterminer la polarité, faire évoluer notre visualisation pour colorer d'une couleur les mots-clés ayant contribué au choix du sentiment prédit et d'une autre couleur les mots-clés ayant contribué au sentiment inverse. Nous pourrions également travailler sur les techniques de résumé automatique, apparié avec des méthodes d'analyse de sentiments basées sur les facettes afin d'améliorer la visualisation.

Nous prévoyons également de travailler sur une interprétation globale du modèle. Tout d'abord, nous pourrions définir la "représentativité" des explications. L'intuition est la suivante. Les bonnes explications sont souvent générales. La représentativité d'une explication correspondrait au nombre de textes couverts par cette explication. Nous pourrions également détecter les explications fréquemment associées via les mots-clés se retrouvant dans les nuages de mots. Ces deux informations pourrait aider l'utilisateur à comprendre sur quoi le classifieur s'appuie généralement pour prendre sa décision. Par opposition, nous pourrions repérer les explications "anormales" (peu fréquentes). Si l'une des caractéristiques d'entrée d'une prédiction est anormale (un groupe de mots-clés rares), elle devrait être présentée dans l'explication du modèle afin d'identifier les cas particuliers. Pour conclure, un outil permettant de naviguer dans les explications basé sur la recherche d'explications représentatives, fréquemment associées et anormales permettrait d'améliorer l'interprétabilité globale du modèle.

IC 2019

7 Remerciement

Nous tenons à remercier la Région Occitanie et la Communauté d'Agglomération de Béziers Méditerranée pour le financement de la thèse de Waleed Ragheb.

Références

ANDERSON P., HE X., BUEHLER C., TENEY D., JOHNSON M., GOULD S. & ZHANG L. (2017). Bottom-up and top-down attention for image captioning and VQA. *CoRR*, **abs/1707.07998**, 6077–6086.

BAHDANAU D., CHO K. & BENGIO Y. (2014). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.

DEVLIN J., CHANG M., LEE K. & TOUTANOVA K. (2018). BERT : pre-training of deep bidirectional transformers for language understanding. *CoRR*, **abs/1810.04805**.

DU J., GUI L., HE Y. & XU R. (2017). A convolutional attentional neural network for sentiment classification. In *International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, p. 445–450.

HERLOCKER J. L., KONSTAN J. A. & RIEDL J. (2000). Explaining collaborative filtering recommendations. In *ACM Conference on Computer Supported Cooperative Work*, CSCW '00, p. 241–250, New York, NY, USA : ACM.

HOWARD J. & RUDER S. (2018). Universal language model fine-tuning for text classification. In *56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 328–339.

HU L., CHEN J., NAIR V. N. & SUDJIANTO A. (2018). Locally interpretable models and effects based on supervised partitioning (LIME-SUP). *CoRR*, **abs/1806.00663**.

LETARTE G., PARADIS F., GIGUÈRE P. & LAVIOLETTE F. (2018). Importance of self-attention for sentiment analysis. In *EMNLP Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, p. 267–275.

LI X., BING L., LI P., LAM W. & YANG Z. (2018). Aspect term extraction with history attention and selective transformation. *CoRR*, **abs/1805.00760**.

LIN Z., FENG M., DOS SANTOS C. N., YU M., XIANG B., ZHOU B. & BENGIO Y. (2017). A structured self-attentive sentence embedding. In *International Conference on Learning Representations (ICLR)*.

LUO L., AO X., PAN F., WANG J., ZHAO T., YU N. & HE Q. (2018). Beyond polarity : Interpretable financial sentiment analysis with hierarchical query-driven attention. In *27th International Joint Conference on Artificial Intelligence, IJCAI*, p. 4244–4250.

MA Y., PENG H. & CAMBRIA E. (2018). Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. In *AAAI Conference on Artificial Intelligence*, p. 5876–5883.

MAAS A. L., DALY R. E., PHAM P. T., HUANG D., NG A. Y. & POTTS C. (2011). Learning word vectors for sentiment analysis. In *49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies - Volume 1*, HLT '11, p. 142–150, Stroudsburg, PA, USA : Association for Computational Linguistics.

MÄNTYLÄ M. V., GRAZIOTIN D. & KUUTILA M. (2018). The evolution of sentiment analysis - a review of research topics, venues, and top cited papers. *Computer Science Review*, **27**, 16–32.

MERITY S., KESKAR N. S. & SOCHER R. (2018). Regularizing and optimizing LSTM language models. In *International Conference on Learning Representations (ICLR)*.

MERITY S., XIONG C., BRADBURY J. & SOCHER R. (2016). Pointer sentinel mixture models. *CoRR*, **abs/1609.07843**.

MOHAMMAD S. M. (2018). Word affect intensities. In *11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

MOHAMMAD S. M. & TURNEY P. D. (2013). Crowdsourcing a word-emotion association lexicon. In *Computational Intelligence*, volume 29, p. 436–465.

MOKHTARI S., LI T. & XIE N. (2018). Context-sensitive neural sentiment classification. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, p. 293–299.

MOLNAR C. (2019). *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

PETERS M., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of*

the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers), p. 2227–2237, New Orleans, Louisiana : Association for Computational Linguistics.

RIBEIRO M. T., SINGH S. & GUESTRIN C. (2016). "why should i trust you ?" : Explaining the predictions of any classifier. In *22nd ACM International Conference on Knowledge Discovery and Data Mining SIGKDD*, p. 1135–1144, New York, NY, USA : ACM.

SU J., ZENG J., XIONG D., LIU Y., WANG M. & XIE J. (2018). A hierarchy-to-sequence attentional neural machine translation model. In *IEEE/ACM Trans. Audio, Speech & Language Processing*, volume 26, p. 623–632.

TUTEK M. & ŠNAJDER J. (2018). Iterative recursive attention model for interpretable sequence classification. In *EMNLP Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, p. 249–257.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30*, p. 5998–6008. Curran Associates, Inc.

VELLIDO A., MARTÍN-GUERRERO J. D. & LISBOA P. J. G. (2012). Making machine learning models interpretable. In *European Symposium on Artificial Neural networks, computational intelligence and machine learning*.

VOULODIMOS A., DOULAMIS N., DOULAMIS A. & PROTOPAPADAKIS E. (2018). Deep learning for computer vision : A brief review. In *Computational Intelligence and Neuroscience*, volume 2018, p. 1–13.

WAN L., ZEILER M., ZHANG S., CUN Y. L. & FERGUS R. (2013). Regularization of neural networks using dropconnect. In *30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, p. 1058–1066, Atlanta, Georgia, USA : PMLR.

WANG S., HUANG M. & DENG Z. (2018). Densely connected cnn with multi-scale feature attention for text classification. In *IJCAI*.

YANG Z., YANG D., DYER C., HE X., SMOLA A. & HOVY E. (2016). Hierarchical attention networks for document classification. In *Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1480–1489 : Association for Computational Linguistics.

YOUNG T., HAZARIKA D., PORIA S. & CAMBRIA E. (2018). Recent trends in deep learning based natural language processing [review article]. In *IEEE Computational Intelligence Magazine*, volume 13, p. 55–75.

ZHANG X., ZHAO J. & LECUN Y. (2015). Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28*, p. 649–657.

Apprentissage de préférences à partir d'une ontologie formelle : méthodes et application en antibiothérapie⋆

Jean-Baptiste Lamy[1], Karima Sedki[1], Rosy Tsopra[2]

[1] LIMICS, Université Paris 13, 93017 Bobigny, France, INSERM UMRS 1142, Sorbonne Universités, Paris
jean-baptiste.lamy@univ-paris13.fr, karima.sedki@univ-paris13.fr

[2] Université Paris 13, SMBH, Bobigny, France ; AP-HP, Assistance Publique des Hôpitaux de Paris, Paris, France
rosy.tsopra@nhs.net

Résumé : L'apprentissage des préférences est un problème de recherche qui a reçu beaucoup d'attention en intelligence artificielle ces dernières années. Il s'agit d'apprendre un modèle de préférences à partir de préférences observées. Ce modèle peut ensuite être utilisé pour obtenir une meilleure compréhension des préférences, et/ou pour effectuer des prédictions sur de nouvelles instances du problème. Les préférences sont généralement apprises à partir de données, par exemple sous la forme d'une matrice « instances × attributs ». Cependant, lorsque le domaine d'application est complexe, il peut être intéressant d'effectuer l'apprentissage à partir d'une ontologie formelle. Mais cela est plus difficile car (1) tous les attributs impactant les préférences n'ont pas nécessairement le même domaine, et (2) certains attributs peuvent correspondre à des propriétés *n*-aires (réifiées dans l'ontologie).

Dans cet article, nous proposons une méthode pour l'apprentissage d'un modèle de préférence à partir d'une ontologie, et prenant en compte notamment des attributs définis sur des domaines multiples, des propriétés *n*-aires, et des valeurs manquantes. Les préférences apprises peuvent alors venir enrichir l'ontologie. Elle « projette » tous les attributs sur une même classe d'individus, et ensuite des méthodes d'apprentissage de préférences plus classiques peuvent être employées. Nous avons appliqué cette méthode à une ontologie des antibiotiques. À partir de cette ontologie, nous avons explicité le modèle de préférence utilisé par les experts lorsqu'ils recommandent des antibiotiques dans les guides de bonnes pratiques, et en particulier quelle sont les propriétés importantes pour un antibiotique et comment les experts gèrent les informations manquantes.

Mots-clés : Ontologies, Préférences, Apprentissage, Antibiothérapie, Métaheuristique

1 Introduction

De nombreux consommateurs utilisent les sites de e-commerce pour acheter des livres, louer des DVD, choisir une voiture ou une recette de cuisine, *etc*. Afin de faciliter leur choix, des systèmes de recommandation (Adomavicius & Tuzhilin, 2005) sont apparus dans les sites de e-commerce. De manière similaire, dans de nombreux domaines professionnels, les systèmes de recommandation ont été proposés pour aider à la prise de décision, sur la base des préférences des experts. Ces systèmes ont aussi été appliqués en médecine (Sezgin & Ozkan, 2013), pour aider un patient à choisir un médecin (Han *et al.*, 2018) ou pour aider un médecin à choisir un traitement en s'appuyant sur les recommandations d'experts médicaux disponibles dans les guides de bonnes pratiques cliniques (Bouaud & Lamy, 2013). Ces systèmes nécessitent l'acquisition de préférences expertes ou utilisateurs. Celles-ci sont souvent coûteuses et longues à éliciter, notamment si le nombre d'instances (*e.g.* produits) est grand. Dans ce cas, il est intéressant d'apprendre les préférences à partir des données, car les données sont plus faciles à observer et à collecter : c'est l'*apprentissage des préférences* (Fürnkranz & Hüllermeier, 2010), qui est un domaine de recherche qui reçoit beaucoup d'attention ces derniers temps dans des disciplines telles que l'intelligence artificielle, l'aide à

⋆. 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 RaMiPA (AAP 2016).

la décision,... Il s'agit d'apprendre automatiquement un modèle de préférences à partir d'observations concrètes sur les préférences. Une fois le modèle appris, il peut être utilisé pour acquérir une meilleure compréhension du domaine et/ou pour aider à la décision.

Cependant, la qualité du modèle de préférences appris dépend du type de données fournies en entrée au système. De nombreuses approches d'apprentissage de préférences n'attachent que peu d'importance à la structure des données, et il s'agit souvent d'instances décrites par un ensemble d'attributs, par exemple dans une simple matrice « instance $\times$ attribut » à deux dimensions. L'utilisation d'ontologies pour structurer les données a été récemment proposée dans les systèmes de recommandations, et il a été montré que cela améliore les performances de ces systèmes (Bagherifard *et al.*, 2017; Werner *et al.*, 2014; Subramaniyaswamy *et al.*, 2013). Les ontologies ont aussi été utilisées pour structurer les modèles de préférences (Ngoc *et al.*, 2005). Cependant, l'apprentissage de préférences à partir d'ontologies est plus complexe, car tous les attributs impactant les préférences ne correspondent pas nécessairement à des propriétés définies sur le même domaine, et certains attributs peuvent correspondre à des propriétés *n*-aires, réifiées en plusieurs propriétés binaires dans l'ontologie.

Dans cet article, nous proposons une méthode pour apprendre un modèle de préférences à partir d'une ontologie, et prenant en compte les domaines multiples, les propriétés *n*-aires, et les valeurs manquantes. La méthode "projette" tous les attributs sur les individus d'une seule classe de l'ontologie, et ensuite les méthodes classiques d'apprentissage de préférences peuvent être employées. Nous décrivons aussi une méthode simple d'apprentissage de préférences, qui traduit l'apprentissage du modèle de préférences en un problème d'optimisation qui est résolu à l'aide d'une métaheuristique. Cependant, d'autres méthodes d'apprentissage pourraient être utilisées, et donc, la méthode que nous proposons pour appliquer l'apprentissage de préférences aux ontologies est générale, et non limitée à la méthode d'apprentissage présentée ici. Nous présentons également une étude de cas, dans laquelle nous utilisons la méthode proposée sur une ontologie des antibiotiques. Cette ontologie décrit les antibiotiques, les maladies infectieuses et les profils de patients concernés. Elle inclut aussi les recommandations de prescription issues de guides de bonnes pratiques cliniques et diverses propriétés liées aux antibiotiques et à leur utilisation (comme la présence d'effets indésirables importants, ou le risque de résistances bactériennes). A partir de cette ontologie, nous rendrons explicite le modèle de préférences implicite utilisé par les experts pour recommander un antibiotique lors de l'écriture des recommandations.

Dans un précédent travail (Tsopra *et al.*, 2018), nous avions utilisé l'apprentissage des préférences pour détecter des inconsistances dans les guides de bonnes pratiques en antibiothérapie. Ici, nous généraliserons la méthode d'apprentissage des préférences à partir d'ontologies, et nous l'étendrons pour gérer les difficultés mentionnées précédemment (domaines multiples, propriétés *n*-aires). De plus, la méthode publiée précédemment était limitée au classement bipartite (*i.e.* avec seulement deux niveaux de préférences, par exemple recommandé et non recommandé) et à des attributs Booléens. Nous généraliserons la méthode à plusieurs niveaux de préférences, et aux attributs catégoriels (ordonnés ou non). Nous améliorerons également le modèle de préférences et son apprentissage, de sorte à limiter les ambiguïtés dans le modèle et à rendre l'apprentissage plus reproductible. Enfin, nous nous intéresserons aux valeurs manquantes, et nous chercherons à comprendre comment elles influencent les préférences (par exemple comment un expert gère une valeur manquante).

La suite de l'article est organisée comme suit. La section 2 décrit la méthode proposée pour générer une matrice pour l'apprentissage des préférences, à partir d'une ontologie où les attributs d'intérêts sont définis sur plusieurs domaines, dont certains sont des propriétés n-aires, et qui comprend des valeurs manquantes. La section 3 détaille le modèle de préférences que nous proposons, la manière dont l'apprentissage de ce modèle peut se traduire en un problème d'optimisation, et comment nous avons résolu ce problème. La section 4 présente une étude de cas utilisant la méthode proposée sur une ontologie des antibiotiques. Enfin, la section 5 discute la méthode proposée avant de conclure en section 6.

2 Générer une matrice simple à partir d'une ontologie

Avant l'apprentissage des préférences proprement dit, il est nécessaire d'extraire à partir de l'ontologie un jeu de données structuré selon un format plus simple, c'est-à-dire une matrice « instance × attribut ». Cette tâche n'est pas triviale, car les attributs d'intérêt pour l'apprentissage des préférences ne sont pas nécessairement tous définis sur le même domaine, et certains de ces attributs peuvent être des propriétés n-aires (réifiées dans l'ontologie).

Considérons une ontologie $\mathcal{O}$ dont les axiomes décrivent un ensemble d'individus $\mathcal{I}$, un ensemble de classes $\mathcal{C}$ et un ensemble de propriétés $\mathcal{R}$. Une classe particulière d'individus $\mathcal{X} \equiv \{x_1, ...\} \subseteq \mathcal{I}$ représente les *instances* pour le besoin de l'apprentissage des préférences. De plus, un sous-ensemble des propriétés $\mathcal{F} = \{p_1, ...\} \subseteq \mathcal{R}$ représente les attributs pris en compte par l'apprentissage des préférences. Les valeurs des attributs peuvent être assertées à différents niveaux dans l'ontologie : sur une instance, *e.g.* $p(x, v)$ où v est la valeur, mais aussi sur une classe d'instance, *e.g.* $c \sqsubseteq \exists p.\{v\}$ avec $c \sqsubseteq \mathcal{X}$, et même sur un individu ou une classe non-instance du problème d'apprentissage de préférences, *e.g.* $p(i, v)$ avec $i \in \mathcal{I} \setminus \mathcal{X}$ et $c \sqsubseteq \exists p.\{v\}$ avec $c \in \mathcal{C} \sqcap \neg\mathcal{X}$. Les attributs peuvent donc avoir n'importe quel domaine (pas nécessairement les instances). Enfin, nous considérons un ensemble de formules de préférences $\mathcal{P}$ observées entre les instances, chaque formule définissant un ordre partiel sur $\mathcal{X}$.

Example 1 : Une ontologie pour un site de *e*-commerce contient les classes suivantes : $Produit$, $Fabricant$ et $Pays$. Les attributs sont $estEnPromotion$ (domaine : $Produit$, range : Booléen) et $aPourPays$ (domaine : $Fabricant$, range : $Pays$). Une propriété non-attribut (vis-à-vis de l'apprentissage de préférence) est $fabriquéPar$ (domaine : $Produit$, range : $Fabricant$). Les formules de préférences expriment les préférences observées sur différents clients, *e.g.* $x_1 \succ x_2 \approx x_3$ (*i.e.* x_1 est préféré à x_2 et x_3 ; mais x_2 et x_3 sont indifférents et aucun n'est préféré à l'autre) si un client a regardé les produits x_1, x_2 et x_3, et a finalement acheté le produit x_1. L'objectif est de comprendre les raisons pour lesquelles un produit est préféré à un autre, et de pouvoir suggérer d'autres produits aux clients, sur la base des deux attributs (promotion et pays de fabrication). Ici, les individus de la classe $Produit$ sont les instances pour l'apprentissage des préférences. Cependant, les deux attributs ne sont pas définis sur le même domaine, alors que l'apprentissage des préférences considère généralement que les attributs portent tous sur les instances.

L'ontologie peut être utilisée pour « projeter » chaque attribut sur les instances, afin de produire cette matrice à partir d'une ontologie complexe avec des attributs de domaines hétérogènes.

2.1 Attributs définis sur un domaine non-instance

Pour chaque attribut p dont le domaine n'est pas les instances du problème d'apprentissage de préférences (*i.e.* $\exists p.\top \not\sqsubseteq \mathcal{X}$), nous définissons une composition de propriétés $q \circ ... \circ p$ qui commence par q (dont le domaine est $\mathcal{X}$, les instances), et qui se termine par p. Nous créons une nouvelle propriété p' avec pour domaine $\mathcal{X}$ et le même range que p. Ensuite, pour chaque valeur possible v du range de p, nous créons une classe c_v équivalente à toutes les instances indirectement reliées à v, et nous assertons une relation directe selon p'. D'un point de vue formel, $c_v \equiv \mathcal{X} \sqcap q \circ ... \circ p.\{v\}$ et $c_v \sqsubseteq \exists p'.\{v\}$. De plus, les valeurs de l'attribut p peuvent être assertées au niveau des classes et non des individus. Pour chaque classe $Y \sqsubseteq \exists p.V$ (où V est la classe de valeur), nous créons une classe C_V avec $c_V \equiv \mathcal{X} \sqcap q \circ ... \circ p.V$ et $c_V \sqsubseteq \exists p'.V$.

Example 1 (suite) : Nous créons la propriété $aPourPaysDeFabrication$ (domaine : $Produit$, range : $Pays$). Ensuite, nous définissons la classe de tous les produits fabriqués par un fabricant en France et nous assertons qu'ils sont fabriqués en France :

$$\begin{aligned} ProduitFaitEnFrance &\equiv Produit \\ &\sqcap \exists fabriquéPar \circ aPourPays.\{France\} \\ ProduitFaitEnFrance &\sqsubseteq \exists aPourPaysDeFabrication.\{France\} \end{aligned}$$

IC 2019

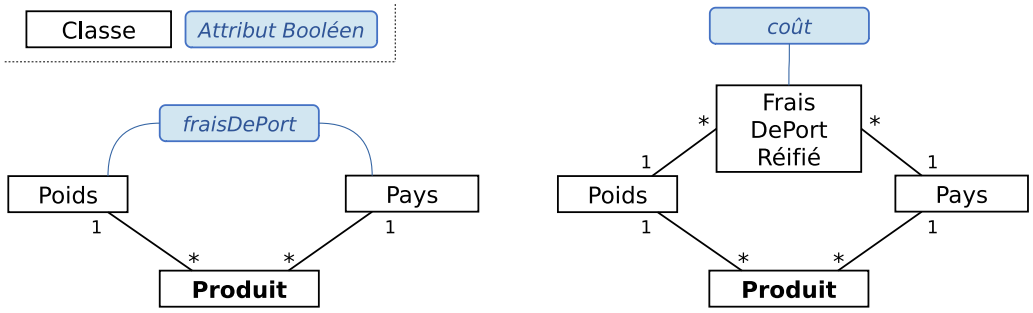


FIGURE 1 – Exemple d'attribut n-aire, avant réification (à gauche) et après (à droite).

2.2 Attributs n-aires

Pour chaque attribut n-aire, nous considérons sa réification en n propriétés binaires, p_1 à p_n. Nous distinguons l'une des entités reliées comme étant le range, c'est-à-dire la cible pour l'apprentissage des préférences, et les $n-1$ autres seront vues comme le domaine. Nous dénoterons arbitrairement le range avec l'index 1. Comme précédemment, nous créons une nouvelle propriété p', avec pour domaine $\mathcal{X}$ et le même range que p_1. Pour chaque valeur possible $(v_1, ..., v_n)$ pour les propriétés $(p_1, ..., p_n)$, nous créons une classe $c_{v_2,...,v_n}$, définie comme l'intersection des instances reliées à v_2,...,v_n *via* les propriétés p_2,...,p_n, et nous assertons une relation directe selon p'. D'un point de vue formel, $c_{v_2,...,v_n} \equiv \mathcal{X} \sqcap p_2.\{v_2\} \sqcap ... \sqcap p_n.\{v_n\}$ et $c_{v_2,...,v_n} \sqsubseteq \exists p'.\{v_1\}$. De plus, pour chaque classe $Y \sqsubseteq \exists p_1.V_1 \sqcap ... \sqcap p_n.V_n$, nous créons une classe $c_{V_2,...,V_n}$ avec $c_{V_2,...,V_n} \equiv \mathcal{X} \sqcap p_2.V_2 \sqcap ... \sqcap p_n.V_n$ et $c_{V_2,...,V_n} \sqsubseteq \exists p'.V_1$.

Exemple 2 : Nous pouvons étendre l'exemple précédent avec les frais de port, qui peuvent impacter la décision du client. Pour rester simple, nous considérerons seulement deux niveaux de frais : $Vrai$ (*i.e.* payant) et $Faux$ (frais de port gratuit). Les frais de port dépendent du poids du produit, avec deux poids possibles : $Léger$ et $Lourd$, et le pays de fabrication (en considérant que le fabricant envoie directement le produit au client). Les frais de port sont donc un attribut ternaire entre le coût, le poids et le pays (Figure 1). Le coût est le range pour l'apprentissage des préférences, tandis que le poids et le pays sont des domaines. Par exemple, la relation $fraisDePort(Faux, Léger, France)$ signifie que les produits légers fabriqués en France sont envoyés gratuitement. Cette relation ternaire peut être « projetée » sur les produits comme suit :

$$\begin{aligned} ProduitLégerFrançais &\equiv Produit \\ &\sqcap \exists aPourPoids.\{Léger\} \\ &\sqcap \exists aPourPaysDeFabrication.\{France\} \\ ProduitLégerFrançais &\sqsubseteq \exists aPourFraisDePort.\{Faux\} \end{aligned}$$

Quand le nombre d'instances de l'attribut n-aire est élevé, la création manuelle de ces classes peut être longue. Cependant, cela peut être automatisé avec un programme qui « pré-processe » l'ontologie. Nous avons utilisé des scripts Python avec Owlready 2 (Lamy JB, 2016, 2017), un module Open Source pour la programmation orientée ontologie.

Les deux difficultés traitées ci-dessus (domaines multiples et attributs n-aires) peuvent être rencontrées sur le même attribut (*i.e.* un attribut n-aire dont le domaine n'est pas les instances de l'apprentissage de préférences). Dans ce cas, les deux solutions proposées peuvent être utilisées conjointement.

Enfin, à l'aide d'un raisonneur comme HermiT (Motik *et al.*, 2009), nous reclassons les instances selon les classes définies par intention (c_v, c_V, $c_{v_2,...,v_n}$ et $c_{V_2,...,V_n}$), et nous obtenons donc pour chaque instance les valeurs des propriétés p'. Pour une instance et un attribut donné, si aucune valeur n'est obtenue, cela correspond à une valeur manquante (*e.g.* si le pays

de fabrication est inconnu). Si plus d'une valeur est obtenue, cela correspond à des valeurs conflictuelles. Dans ce cas, en fonction de l'attribut, il est possible de : (a) garder la plus mauvaise valeur, si les valeurs sont ordonnées, ou (b) conserver toutes les valeurs conflictuelles. La méthode d'apprentissage de préférences que nous décrirons par la suite autorise les deux options.

3 Apprentissage des préférences

L'apprentissage des préférences est réalisé sur les instances $\mathcal{X}$ et les attributs $\mathcal{F}' = \{p'_1, ...\}$, où p'_i sont les nouvelles propriétés créées à la section précédente, ou à défaut $p'_i = p_i$ pour les attributs binaires qui ont pour domaine les instances (pour lesquels aucune nouvelle propriété n'est nécessaire). Pour un attribut p', nous notons $\mathcal{V}_{p'}$ l'ensemble des valeurs possibles.

3.1 Modèle de préférences

De nombreuses méthodes d'apprentissage de préférences ont été proposées dans la littérature (Koriche & Zanuttini, 2010; Dekel *et al.*, 2003; Burges *et al.*, 2005). Celle que nous présentons ici vise à apprendre à la fois :

— **des contraintes nécessaires simples** $\mathcal{N}$, de la forme $p' = v$,
— **des préférences** $\mathcal{W}$, exprimées par des poids associés à chaque valeur possible de chaque attribut.

Les contraintes correspondent à des critères que doivent obligatoirement satisfaire les instances : par exemple, seuls les produits disponibles dans le pays du client peuvent être proposés à la vente. Au contraire, les préférences quantifient l'importance des attributs et de leurs valeurs. Nous pouvons donc formaliser notre modèle de préférences par un couple $\mathcal{M} = (\mathcal{N}, \mathcal{W})$ où $\mathcal{N}$ est un sous-ensemble de l'ensemble des contraintes possibles et $\mathcal{W}$ est une liste de poids, avec un poids pour chaque valeur possible pour chaque attribut :

$$\mathcal{M} = \begin{cases} \mathcal{N} \subseteq \{p' = v : \forall p' \in \mathcal{F}', \forall v \in \mathcal{V}_{p'}\} \\ \mathcal{W} = \left(w_{p',v} : \forall p' \in \mathcal{F}', \forall v \in \mathcal{V}_{p'}\right) \end{cases}$$

Notons que, si la contrainte $p' = v$ est présente dans $\mathcal{N}$, tous les poids $w_{p',v'}$ pour la propriété p' ne seront pas utilisés par la suite. Cependant, nous laissons ces poids présents dans le modèle pour faciliter l'apprentissage, car l'apprentissage des contraintes et des poids se fait simultanément.

Les valeurs manquantes sont associées à un poids de 0. Il s'agit d'une « origine arbitraire » pour la mesure des poids : comme les poids n'ont pas d'unité ni d'origine, nous pouvons définir cette origine à notre guise. Pour les valeurs conflictuelles, la somme des poids des différentes valeurs est considérée.

3.2 Réduction à un problème d'optimisation

Tout d'abord, pour les instances satisfaisant les contraintes nécessaires, nous définissons une fonction d'utilité u qui calcule son utilité. Les instances avec une plus grande utilité sont préférées aux autres. Nous avons choisi ici une fonction u très simple qui calcule la somme des poids pour chaque valeur d'attribut associée à l'instance :

$$u(x_i) = \sum \{w_{p',v} \in \mathcal{W} : p'(x_i, v)\}$$

Ensuite, nous définissons la fonction E (Algorithm 1) qui calcule le taux d'erreur obtenu lorsque l'on compare le modèle $\mathcal{M}$ aux formules de préférences observées $\mathcal{P}$ sur les instances $\mathcal{X}$. E calcule d'abord l'ordre total $\mathcal{T}$ sur $\mathcal{X}$, en utilisant le modèle : les instances qui satisfont les contraintes nécessaires sont préférées aux autres, et, parmi celles qui satisfont les

Algorithme 1 Algorithme de la fonction E qui calcule le taux d'erreur du modèle.

fonction $E(\mathcal{X}, \mathcal{P}, \mathcal{M})$:
soit $\mathcal{N}$ et $\mathcal{W}$ les deux composantes du modèle : $\mathcal{M} = (\mathcal{N}, \mathcal{W})$
soit $\mathcal{X}_{\mathcal{N}} = \{x \in \mathcal{X} : x \text{ satisfait les contraintes nécessaires } \mathcal{N}\}$
soit $\mathcal{X}_{\overline{\mathcal{N}}} = \mathcal{X} \setminus \mathcal{X}_{\mathcal{N}}$
soit $e = 0$ le nombre d'erreurs trouvés
soit $\mathcal{T}$ un ordre total sur $\mathcal{X}$, défini comme suit :
$\forall x_i \in \mathcal{X}_{\mathcal{N}}, \forall x_j \in \mathcal{X}_{\mathcal{N}}, u(x_i) > u(x_j) \Leftrightarrow x_i \succ x_j$
$\forall x_i \in \mathcal{X}_{\mathcal{N}}, \forall x_j \in \mathcal{X}_{\overline{\mathcal{N}}},\ x_i \succ x_j$
$\forall x_i \in \mathcal{X}_{\overline{\mathcal{N}}}, \forall x_j \in \mathcal{X}_{\overline{\mathcal{N}}},\ x_i \approx x_j$
pour chaque instance $x \in \mathcal{X}$:
pour chaque formule de préférence $p \in \mathcal{P}$ impliquant x :
si p n'est pas compatible avec l'ordre total $\mathcal{T}$:
$e = e + 1$
break
retourne $\frac{e}{|\mathcal{X}|}$

contraintes, celles qui ont une plus grande utilité sont préférées. Enfin, la fonction compare $\mathcal{T}$ avec les formules de préférences pour chaque instance, et retourne le taux d'erreur E.

Afin d'augmenter la clarté du modèle (pour un humain), la capacité à le traduire visuellement par la suite, et la reproductibilité de l'apprentissage, nous avons ajouté deux contraintes au problème d'optimisation, détaillées dans les deux paragraphes suivants.

Premièrement, nous avons utilisé des valeurs entières pour les poids. Lors d'un précédent travail (Tsopra *et al.*, 2018), nous avions utilisé des poids réels. Cependant, cela conduisait à une certaine ambiguïté dans le modèle. Par exemple un poids appris de 0,71 pour un attribut doit-il être considéré comme significativement différent d'un poids de 0,73 pour un autre attribut ? Ou bien ces deux poids doivent-ils être considérés comme équivalents ? Dans la mesure où l'apprentissage n'est pas reproductible, une autre exécution peut éventuellement aboutir à des poids différents pour ces deux propriétés. Afin d'éviter ces ambiguïtés, dans le présent travail nous avons restreint les poids à des valeurs entières. De plus, afin de faciliter l'affichage graphique par la suite, nous avons interdit les valeurs 1 et -1 pour les poids. En effet, si l'on traduit visuellement les poids par la hauteur d'un rectangle, les valeurs trop petites conduiront à des rectangles de faible hauteur, dans lesquels il sera difficile d'écrire le nom de l'attribut correspondant. Par conséquent, $w_{p',v} \in \{..., -4, -3, -2, 0, 2, 3, 4...\}$.

Deuxièmement, lors de l'apprentissage, nous cherchons à minimiser le taux d'erreur, mais aussi la somme totale des poids, notée S_w. Notons que, si la contrainte $p' = v$ est présente dans $\mathcal{N}$, tous les poids $w_{p',v'}$ n'ont aucun impact. Par conséquent, nous calculons la somme des poids en excluant ceux qui, du fait des contraintes nécessaires, n'ont aucun impact, de la manière suivante :

$$S_w = \sum_{p'} \sum_{v'} \{w_{p',v} : \nexists v' \, such \, as \, (p' = v') \in \mathcal{N}\}$$

Minimiser la somme des poids présente plusieurs intérêts. Tout d'abord, en multipliant tous les poids par une valeur constante n, il est possible d'obtenir un modèle différent mais conduisant au même taux d'erreur. Ici, nous évitons ce problème et augmentons donc la reproductibilité de l'apprentissage. De plus, comme les poids ne sont pas comptés lorsqu'il existe une contrainte nécessaire, cela biaise l'apprentissage en faveur des contraintes : si deux modèles ont le même taux d'erreur mais que l'un contient davantage de contraintes nécessaires, il sera préféré. Cela est intéressant car les contraintes sont plus faciles que les poids à appréhender pour un humain. Enfin, si les poids sont présentés visuellement, par exemple par des hauteurs, cela permet de réduire l'espace occupé par la visualisation et de la rendre plus compacte.

Par conséquent, nous cherchons le modèle $\mathcal{M}^{best}$ qui minimise à la fois le taux d'erreur E et la somme des poids S_w, selon un ordre lexicographique (*i.e.* nous minimisons le taux

Algorithme 2 Algorithme pour les fonctions vol et $marche$ utilisées pour l'optimisation.

fonction $vol()$:
  soit $\mathcal{N}$ un ensemble de contraintes aléatoires
  soit $\mathcal{W} = \big(w_{p',v} =$ nombre entier différent de -1 et $1 : \forall p' \in \mathcal{F}', \forall v \in \mathcal{V}_{p'}\big)$
  retourne $\mathcal{M} = (\mathcal{N}, \mathcal{W})$

fonction $marche(\mathcal{M})$:
  soit $\mathcal{M}' = (\mathcal{N}', \mathcal{W}')$ une copie de $\mathcal{M}$
  soit r un nombre réel aléatoire entre 0 et 1
  si $r < 0.15$:
    ajouter une contrainte aléatoire à $\mathcal{N}'$
  sinon si $r < 0.3$:
    enlever une contrainte aléatoire de $\mathcal{N}'$
  sinon :
    modifier un poids de $\mathcal{W}'$ en utilisant la fonction de $marche$ proposée
    pour résoudre les problèmes d'optimisation global non-linéaire
    dans la métaheuristique AFB (Lamy JB, 2019)
  retourne $\mathcal{M}'$

d'erreur et, à taux d'erreur égal, nous préférons le modèle ayant la somme des poids la plus faible). Il s'agit d'un problème d'optimisation :

$$\mathcal{M}^{best} = \arg\min_{\mathcal{M}} \big(E(\mathcal{X}, \mathcal{P}, \mathcal{M} = (\mathcal{N}, \mathcal{W})), S_w\big)$$

3.3 Résolution du problème d'optimisation

Le problème d'optimisation décrit à la section précédente est complexe car il mélange optimisation combinatoire (pour optimiser $\mathcal{N}$) et optimisation globale non-linéaire (pour optimiser $\mathcal{W}$). Cependant, $\mathcal{N}$ et $\mathcal{W}$ doivent être optimisés simultanément, puisqu'ils sont interdépendants.

Nous avons résolu ce problème à l'aide de la métaheuristique des Oiseaux Picorant Artificial (OPA, *Artificial Feeding Birds* AFB) (Lamy JB, 2019), inspirée par le comportement des pigeons. Cette métaheuristique est très générique et peut résoudre n'importe quel problème d'optimisation défini par un triplet de trois fonctions $(coût, vol, marche)$, où $coût$ est la fonction de coût à minimiser, vol est une fonction qui retourne une solution aléatoire et $marche$ est une fonction qui modifie légèrement une solution existante. L'algorithme 2 décrit les fonctions vol et $marche$ que nous proposons pour optimiser le modèle de préférences $\mathcal{M}$.

4 Application à une ontologie de l'antibiothérapie

4.1 Contexte

Pour aider les médecins à prescrire le bon antibiotique, les autorités de santé publient des guides de bonnes pratiques cliniques. Dans ces guides, les experts recommandent la prescription de certains antibiotiques, selon les propriétés de chaque médicament, la maladie infectieuse et le profil du patient (adulte, enfant, femme enceinte,...). Par exemple, ils recommandent la fosfomycine-trométamol pour une cystite non compliquée chez la femme. Dans cette étude, nous cherchons à établir le modèle de préférences implicites qu'utilisent les experts pour choisir un antibiotique, lorsque la prescription d'antibiotique est justifiée.

Au cours de travaux précédents (Tsopra *et al.*, 2014a,b), nous avons construit une base de connaissances décrivant 50 antibiotiques dans 66 situations cliniques, selon 11 attributs utilisés par les experts pour établir les recommandations (Table 1). Chaque attribut est Booléen, et sa valeur dépend de l'antibiotique mais également du profil patient, de la maladie infectieuse

#	**Attribut [**$nom\ court$**]** Définition
1	**Naturellement actif contre la bactérie pathogène [**$naturallement\,actif$**]** Est-ce que la bactérie pathogène est décrite comme naturellement sensible à l'antibiotique ? (*e.g.* l'amoxicilline est naturellement active contre les streptocoques du groupe A)
2	**Probablement actif contre la bactérie pathogène [**$probablement\,actif$**]** Est-ce que la fréquence de résistance de la bactérie pathogène pour l'antibiotique est décrite comme suffisamment basse pour permettre sa prescription ? (*e.g.* la ceftriaxone est probablement active contre E.coli)
3	**Efficacité clinique prouvée [**$prouvée$**]** Est-ce que l'antibiotique est décrit comme cliniquement efficace pour traiter l'infection OU est (ou a été) indiqué ou recommandé pour cette infection ?
4	**Absence de contre-indication [**$non\,contre\,indiqué$**]** Est-ce qu'il n'y a pas de contre-indication absolue à la prise de l'antibiotique pour le profil de patient concerné ?
5	**Protocole pratique [**$protocole$**]** Est ce que le protocole d'administration de l'antibiotique est décrit comme pratique pour le patient, c'est-à-dire administration par voie orale ET durée de traitement courte ?
6	**Classe non-précieuse [**$non\,précieux$**]** Est-ce que l'antibiotique appartient à une classe décrite comme « non précieuse », c'est-à-dire qui ne doit pas être préservée pour des infections plus graves ?
7	**Absence d'effets indésirables graves ou fréquents [**$peu\,d'effet\,ind$**]** Est-ce que l'antibiotique est décrit comme ne causant aucun effet indésirable grave et peu d'effets indésirables non-graves ?
8	**Haut niveau d'efficacité [**$très\,efficace$**]** Est-ce que l'antibiotique est décrit comme très efficace pour ce type d'infection ?
9	**Spectre antibactérien étroit [**$spectre$**]** Est-ce que l'antibiotique est décrit comme ayant un spectre antibactérien étroit ?
10	**Faible impact écologique sur les résistances bactériennes [**$risque\,éco\,bas$**]** Est-ce que l'antibiotique est décrit comme ayant un risque faible de provoquer l'émergence de nouvelle résistance bactérienne ?
11	**Goût [**$goût$**]** Est-ce que l'antibiotique a un goût acceptable pour le profil de patient (enfants notamment) ?

TABLE 1 – Les 11 attributs décrivant les antibiotiques dans l'ontologie.

(*e.g.* cystite) et/ou de la bactérie pathogène (*e.g.* Escherichia coli). Ses attributs peuvent donc être des attributs n-aires. La base de connaissances a été construite et peuplée par un médecin (RT) à partir de plusieurs guides de bonnes pratiques cliniques, et ensuite validée par un panel d'experts en antibiothérapie au cours d'un processus Delphi.

Cette base de connaissances a ensuite été formalisée dans une ontologie OWL 2.0 (Tsopra *et al.*, 2018). Elle contient 144,038 triplets RDF décrivant 5.696 classes, 19 propriétés et 34.483 axiomes, et appartenant à la famille $\mathcal{ALC}(\mathcal{D})$ [1] des logiques de description. La Fi-

1. $\mathcal{AL}$: langage d'attributs (incluant la négation atomique, l'intersection de concepts, la restriction univer-

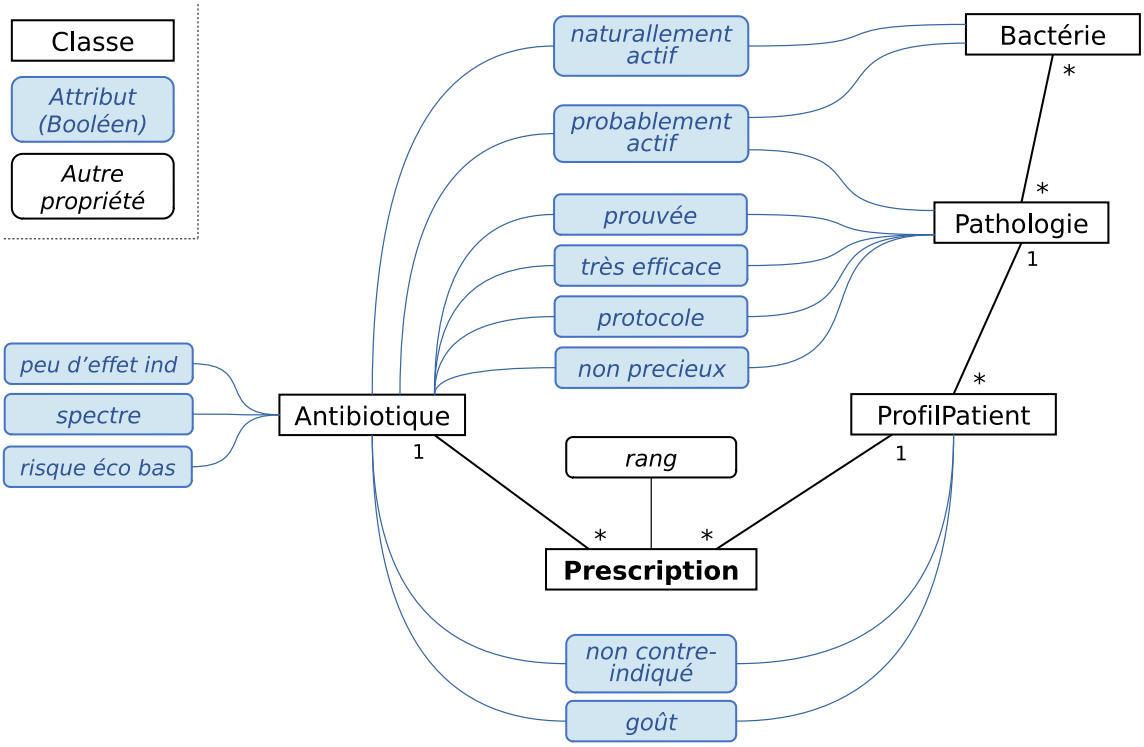


FIGURE 2 – Structure générale de l'ontologie de l'antibiothérapie. Les attributs ternaires et quaternaires (tous sauf les trois sur la gauche) sont réifiés dans l'ontologie.

gure 2 montre la structure générale de l'ontologie. Elle comprend cinq classes principales : $Antibiotique$, $ProfilPatient$ (*e.g.* adulte, enfant ou femme enceinte) associés à une maladie infectieuse ($Pathologie$) causée par une $Bactérie$ pathogène. Une $Prescription$ est l'association d'une $Antibiotique$ avec un $ProfilPatient$ (lui-même relié à une $Pathologie$ et une $Bactérie$). Les 11 attributs sont définis sur 5 domaines différents (aucun d'entre eux n'étant $Prescription$), et incluent 3 propriétés binaires, 7 ternaires et 1 quaternaire. De nombreuses valeurs manquantes sont présentes, à cause de la persistance importante d'information inconnue dans le domaine médical. De plus, pour les $Prescription$ recommandées dans les guides, le rang de recommandation est présent (1 à 4, le rang 1 étant prescrit de préférence au rang 2, *etc*).

Toutes les combinaisons médicaments-situations cliniques ont été créées dans l'ontologie (soit $66 \times 50 = 3300$), sous forme de classes OWL. Celles-ci ont ensuite été reclassées à l'aide du raisonneur HermiT.

4.2 Application de la méthode

Nous avons appliqué la méthode proposée à cette ontologie. $Prescription$ correspond à la classe des instances pour l'apprentissage des préférences. Les rangs de recommandations ont été traduits en formules de préférences $\mathcal{P}$; par exemple si un guide recommande la prescription d'un antibiotique A en rang 1 et la prescription de B ou C en rang 2 pour un $ProfilPatient$ donné, cela a été traduit par la formule de préférence $A \succ B \approx C \succ D \approx E \approx ...$ (où $D, E, ...$ sont les prescriptions correspondant à tous les autres antibiotiques disponibles mais non recommandés par le guide). La Figure 3 montre la structure générale de la matrice obtenue à partir de l'ontologie. Les rangs de recommandations correspondent aux

selle et les qualifieurs existentiels limités à la classe Thing), $\mathcal{C}$: négation complexe, $(\mathcal{D})$: type de données (Baader *et al.*, 2007).

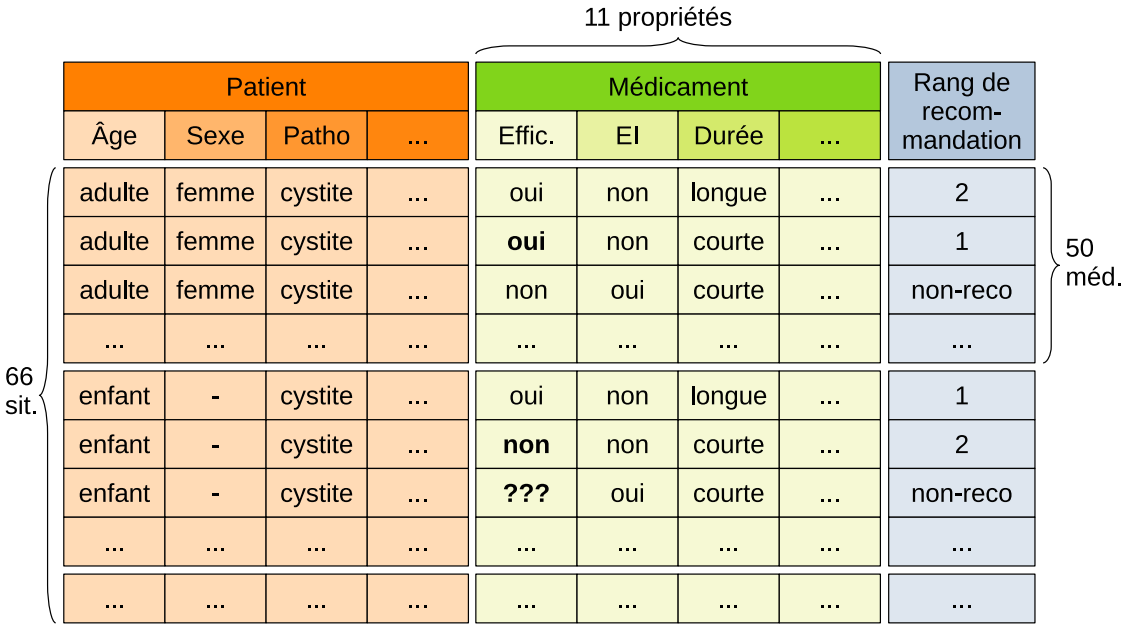


FIGURE 3 – Structure générale de la matrice « instances × attributs » obtenue (EI : effets indésirables).

recommandations trouvées dans les guides de bonnes pratiques cliniques. Pour une situation clinique donnée, la partie « patient » (en orange) reste la même, tandis que la partie « médicament » en vert varie selon les 50 médicaments. Cette partie « médicament » dépend du médicament mais aussi de la partie patient (certaines propriétés du médicament change selon le patient pour lequel il est prescrit, par exemple la durée du traitement n'est pas forcément la même pour un enfant et un adulte).

Lorsque des valeurs conflictuelles ont été rencontrées (*e.g.* une $Pathologie$ associée à deux $Bacteries$ pathogènes putatives, l'une étant résistante à l'antibiotique et l'autre ne l'étant pas), nous avons retenu la pire valeur (principe de précaution, ici la valeur $Faux$). De plus, tous les attributs de l'ontologie correspondent à des avantages potentiels d'un antibiotique (*e.g.* peu d'effets indésirables, grande efficacité, *etc.*). Par conséquent, nous avons restreint les poids associés à la valeur $Faux$ à des nombres négatifs (ou nul), et ceux associés à la valeur $Vrai$ à des nombres positifs (ou nul ; *i.e.* $w_{p',False} \leq 0$ et $w_{p',True} \geq 0$). Ceci interdit d'apprendre un modèle médicalement absurde, par exemple dans lequel les antibiotiques avec beaucoup d'effets indésirables seraient préférés à ceux en ayant peu.

La méthode a été implémentée en Python et l'ontologie a été manipulée avec Owlready (Lamy JB, 2016, 2017). Nous avons utilisé les valeurs par défaut des paramètres pour la métaheuristique AFB (valeurs figurant dans (Lamy JB, 2019)). Nous avons exécuté la métaheuristique sur 3,000 itérations et nous avons retenu le meilleur modèle trouvé.

4.3 Évaluation du processus d'apprentissage

Afin d'évaluer le processus d'apprentissage et la capacité à généraliser les résultats au-delà des données d'apprentissage, nous avons effectué une validation croisée par dixième (*10-fold cross-validation*). La base de données a été découpée en 10 sous-ensembles aléatoires. Pour chaque sous-ensemble, nous avons effectué l'apprentissage sur les 9 autres sous-ensembles, et utilisé le dixième sous-ensemble comme jeu de test. Le taux d'erreur moyen est de 3.5% sur le jeu d'apprentissage et de 5.2% sur le jeu de test. Le taux d'erreur restant faible sur le jeu de test, cela suggère que le modèle de préférences est suffisamment générique pour pouvoir être appliqué à des situations qui ne figurent pas dans le jeu d'apprentissage.

$$\begin{array}{rcl} & naturallement\,actif & = Vrai \\ \wedge & probablement\,actif & = Vrai \\ \wedge & prouvée & = Vrai \\ \wedge & non\,contre\,indicaqué & = Vrai \end{array}$$

	Attribut	$w_{p',Faux}$	$w_{p',Vrai}$
5	$protocole$	-7	4
6	$non\,précieux$	-2	2
7	$peu\,d'effet\,ind$	-5	2
8	$très\,efficace$	-2	2
9	$spectre$	-2	2
10	$risque\,éco\,bas$	-3	2
11	$goût$	-3	2

TABLE 2 – Les contraintes nécessaires $\mathcal{N}$ (en haut) et les poids de préférence $\mathcal{W}$ appris (en bas ; les poids ne figurent pas lorsqu'ils n'ont aucun impact, du fait des contraintes nécessaires).

4.4 Résultats

La Table 2 montre le meilleur modèle obtenu, après apprentissage sur l'ensemble de la base de connaissances. Celui-ci conduit à un taux d'erreur de 3,5%. Le modèle de préférences montre l'importance de chaque attribut pour le choix d'un antibiotique à prescrire.

Les contraintes apprises montrent que les prescriptions doivent avoir nécessairement la valeur $Vrai$ pour quatre attributs ($naturallement\,actif$, $probablement\,actif$, $prouvée$, $non\,contre\,indiqué$) pour être recommandées. Ceci est pertinent d'un point de vue clinique car seuls les antibiotiques actifs d'un point de vue microbiologique et clinique doivent être prescrits, et les contre-indications doivent être respectées. En effet, les attributs impliqués dans les contraintes nécessaires sont ceux considérés comme les plus importants par le médecin de notre équipe (RT). En particulier, trois d'entre eux ($naturallement\,actif$, $prouvée$, $probablement\,actif$) sont liés à l'efficacité du traitement et le troisième ($non\,contre\,indiqué$) à la sécurité.

Parmi les attributs de préférence, l'attribut $protocole$ semble avoir des poids plus importants que les autres. Ceci est logique car le protocole détermine la facilité de prise du médicament, et donc l'observance du patient : les antibiotiques trop compliqués à prendre sont oubliés par le patient. En particulier, si le traitement est trop long, le patient interrompt souvent celui-ci, ce qui conduit à des rechutes et peut favoriser l'émergence de résistances bactériennes.

Ce modèle de préférences permet aussi de comprendre comment les experts traitent les informations manquantes, pour chaque attribut. Les valeurs manquantes correspondent à un poids de 0 (origine arbitraire). Par conséquent, la position du 0 par rapport aux deux bornes w_{Faux} et w_{Vrai} indique si les valeurs manquantes sont plutôt traitées comme des $Faux$ ou des $Vrai$. Ici, dans la Table 2, nous pouvons constater que, pour l'attribut $peu\,d'effet\,ind$, la valeur absolue de $w_{peu\,d'effet\,ind,Faux}$ est beaucoup plus grande que celle de $w_{peu\,d'effet\,ind,Vrai}$. Cela signifie que les valeurs manquantes sont traitées comme étant plus proches de $Vrai$ que de $Faux$, c'est-à-dire que, lorsque l'expert ne sait pas si un antibiotique présente ou non des effets indésirables graves ou fréquents, il aura plutôt tendance à faire comme s'il n'y avait pas de tels effets.

Globalement, nous constatons dans la Table 2 que les valeurs absolues des poids associés aux valeurs $Faux$ sont toujours supérieures ou égales aux poids associés aux valeurs $Vrai$. Les experts interprètent donc les valeurs manquantes comme plutôt « en faveur » du médicament. Cela est particulièrement vrai pour les effets indésirables ($peu\,d'effet\,ind$, comme

vu au paragraphe précédent) et la facilité de prise ($protocole$). En revanche, pour le spectre ($spectre$) et l'efficacité ($très\,efficace$), les valeurs manquantes sont traitées par les experts comme un « entre-deux » situé à mi-chemin entre $Vrai$ et $Faux$.

5 Discussion

Dans cet article, nous avons proposé une méthode générale pour l'apprentissage de préférences à partir d'une ontologie formelle. La méthode permet l'apprentissage de préférences à partir d'attributs qui peuvent être définis sur des domaines hétérogènes, ainsi que d'attributs n-aires. Elle permet aussi d'étudier la manière dont les valeurs manquantes sont traitées. Nous avons aussi décrit un modèle de préférences et une méthode d'apprentissage, fondée sur l'optimisation. Cependant, d'autres méthodes d'apprentissage pourraient être utilisées, telles que l'intégrale de Choquet (Tehrani *et al.*, 2012). Nous avons illustré notre approche sur une ontologie de l'antibiothérapie, et nous avons présenté le modèle de préférences résultant. Comparé à nos travaux antérieurs (Tsopra *et al.*, 2018), ce modèle permet une meilleure compréhension des critères utilisés par les experts pour recommander les antibiotiques, et en particulier de la manière dont ils gèrent les valeurs manquantes. Le modèle présente également moins d'ambiguïtés liées aux valeurs de poids très proches. Les contraintes et les poids appris peuvent ensuite venir enrichir l'ontologie, par exemple sous forme d'annotations associées aux différentes propriétés.

Dans la section 2, notre méthode requiert la création de nombreuses classes et propriétés pour prendre en compte les domaines multiples et les attributs n-aires. Nous avons suggéré l'utilisation d'un langage de programmation pour "prétraiter" l'ontologie et créer ces classes et propriétés. La principale limite de cette approche est qu'elle ne prend en compte que les faits assertés, mais pas les faits qui pourraient être inférés. Une autre solution aurait été l'utilisation de règles fondées sur la logique du premier ordre. Si nous reprenons l'exemple 2 (les frais de port), la règle suivante aurait pu être utilisée :

$$\begin{aligned}\forall(r,w,c,p), & FraisDePortRéifié(r) \wedge aPourValeur(r,Faux)\\ & \wedge Poids(w) \wedge aPourPoids(r,w)\\ & \wedge Pays(c) \wedge aPourPays(r,c)\\ & \wedge Produit(p) \wedge aPourPoids(p,w) \wedge aPourPaysDeFabrication(p,c)\\ & \Rightarrow aPourFraisDePort(p,Faux)\end{aligned}$$

Cette règle est plus générique que la classe $ProduitLégerFrançais$ que nous avons proposé dans l'exemple 2, car elle couvre tous les produits avec des frais de port gratuit, et pas seulement les produits français légers. Cependant, ces règles ont une limitation importante : elles ne fonctionnent que sur les individus, mais pas sur les classes (*e.g.* dans la règle précédente, r, w, c et p doivent être des individus, et non des classes). Par conséquent, l'approche « prétraitement » est préférable si les attributs sont assertés au niveau des classes (sous forme de restrictions), tandis que la logique du premier ordre est préférable si les faits inférés doivent être considérés. Ici, l'ontologie de l'antibiothérapie que nous avons utilisée inclut de nombreux attributs dont les valeurs sont assertés au niveau des classes. Par exemple, toutes les pénicillines A sont contre-indiquées chez les patients allergiques à l'amoxicilline, où « pénicillines A » est une classe d'antibiotiques. C'est pourquoi nous avons choisi l'approche « prétraitement ».

Au contraire, les logiques de description ne peuvent pas être utilisées pour définir des classes générales ciblant chaque valeur d'un attribut n-aire, parce que les logiques de description sont sans variable (*variable-free*) et correspondent en fait à des formules avec une seule variable libre (Baader *et al.*, 2007). Dans l'exemple précédent, il serait tentant de considérer la classe $ProduitÀFraisDePortGratuit$ (une classe correspondant à la règle précédente en logique du premier ordre, plus générale que $ProduitLégerFrançais$), et de la définir de la

sorte :

$$\begin{aligned} ProduitÀFraisDePortGratuit \equiv\ & Produit \\ \sqcap\ & \exists aPourPoids \\ & \quad \circ aPourFraisDePortRéifié \\ & \quad \circ aPourValeur.\{Faux\} \\ \sqcap\ & \exists haPourPaysDeFabrication \\ & \quad \circ aPourFraisDePortRéifié \\ & \quad \circ aPourValeur.\{Faux\} \\ ProduitÀFraisDePortGratuit \sqsubseteq\ & \exists aPourFraisDePort.\{Faux\} \end{aligned}$$

Cependant, cette définition ne fonctionnera pas comme souhaité. Plus précisément, les deux $FraisDePortRéifié$ (celui relié au $Poids$ et celui relié au $Pays$) ne sont pas nécessairement les mêmes. Corriger cela demanderait une seconde variable libre (pour asserter que les deux $FraisDePortRéifié$ sont les mêmes), et cela n'est pas possible en logiques de description.

Dans la littérature, l'apprentissage des préférences a rarement été appliqué aux ontologies. Tsai *et al.* (Tsai *et al.*, 2006) et Wang *et al.* (Wang *et al.*, 2007) ont proposé un modèle d'apprentissage fondé sur une approche ontologique pour les systèmes de e-learning. Il s'appuie sur des cours décrits en SCORM (*Sharable Content Object reference Model*). L'ontologie permet d'inférer les prérequis pour chaque cours, et d'hériter les valeurs des attributs à partir de leur classe. Contrairement au travail que nous avons présenté ici, les auteurs n'ont pas pris en compte les attributs définis sur des domaines différents ni les attributs n-aires. Cependant, l'utilisation d'une sémantique plus riche offerte par l'ontologie formelle peut être utile pour l'apprentissage de préférences dans des domaines complexes, comme l'antibiothérapie.

Nous avons utilisé la métaheuristique AFB, à cause de sa généricité et sa capacité à résoudre des problèmes mélangeant optimisation globale non-linéaire et optimisation combinatoire. En particulier, et contrairement aux autres métaheuristiques existantes (Yang XS, 2010), l'AFB ne nécessite pas le calcul de distance entre solutions du problème. Cela rend l'AFB plus facile à adapter à des problèmes nouveaux pour lesquels le calcul d'une telle distance n'est pas trivial. De plus, nos travaux précédents (Tsopra *et al.*, 2018) ont montré que l'AFB était plus performant pour l'apprentissage de préférences dans ce contexte, comparé à d'autres métaheuristiques.

Nous avons présenté une étude de cas en antibiothérapie. À notre connaissance, nos travaux sont les premiers à essayer d'appliquer l'apprentissage des préférences aux guides de bonnes pratiques cliniques. Dans ce domaine, les valeurs de certains attributs (tels que les résistances des bactéries aux antibiotiques) évoluent rapidement. La méthode que nous proposons est automatique, et donc le modèle de préférences peut être mis à jour facilement lorsque l'ontologie est modifiée.

Les experts qui écrivent les guides s'appuient sur leurs connaissances propres, qui sont essentiellement tacites (Smith MK, 2003). La présentation du modèle de préférences explicite à ces experts pourrait leur permettre de mieux appréhender leurs connaissances et leur processus de raisonnement, et de produire des recommandations plus fiables et plus consistantes.

Dans le domaine thérapeutique, très peu d'approches fondées sur les préférences ont été proposées. La majorité des système d'aide à la décision clinique (Bouaud & Lamy, 2013) s'appuie sur l'implémentation des recommandations des guides et sur les contre-indications des médicaments, *e.g.* sous forme de systèmes à base de règles. Cependant, ils ne cherchent pas à construire un modèle de préférences ni à comprendre les raisons sous-jacentes aux recommandations, comme nous l'avons fait ici. Un système d'aide à la décision s'appuyant sur un modèle de préférences serait donc une approche radicalement nouvelle en médecine. Un tel système permettrait en particulier d'effectuer des prédictions sans se limiter aux recommandations présentes dans les guides. En effet, nous avons vu (section 4.3) que le modèle pouvait être généralisé au-delà des données d'apprentissage. Il pourrait donc servir à produire des recommandations lorsque les recommandations sont absentes (souvent les guides

ne donnent pas des recommandations pour tous les cas possibles (Bouaud *et al.*, 2009)) ou lorsqu'elles ne peuvent pas être appliquées (par exemple à cause de contre-indications ou d'allergie).

6 Conclusion

Nous avons proposé une méthode générale pour l'apprentissage de préférences à partir d'une ontologie formelle, prenant en compte les domaines hétérogènes, les attributs n-aires et les valeurs manquantes. Nous avons montré comment cette approche permettait de mieux comprendre le raisonnement des experts en infectiologie. Les perspectives de ce travail incluent l'application de la méthode à d'autres ontologies dans le domaine médical ou au-delà, ainsi que la visualisation du modèle de préférences appris en antibiothérapie et son utilisation pour l'aide à la décision thérapeutique auprès des médecins généralistes (Tsopra *et al.*, 2019).

Références

ADOMAVICIUS G. & TUZHILIN A. (2005). Toward the next generation of recommender systems : A survey of the state-of-the-art and possible extensions. *IEEE Transactions of Knowledge and Data Engineering*, **17**, 734–749.

BAADER F., CALVANESE D., MCGUINNESS D. L., NARDI D. & PATEL-SCHNEIDER P. L. (2007). *The description logic handbook : theory, implementation and applications*. Cambridge University Press.

BAGHERIFARD K., RAHMANI M., NILASHI M. & RAFE V. (2017). Performance improvement for recommender systems using ontology. *Telematics and Informatics*, **34**(8), 1772–1792.

BOUAUD J. & LAMY J. B. (2013). *Yearbook of medical informatics*, volume 8, chapter A medical informatics perspective on clinical decision support systems. Findings from the yearbook 2013 section on decision support, p. 128–31.

BOUAUD J., SÉROUSSI B., FALCOFF H., JULIEN J., SIMON C. & DENKÉ D L. (2009). Consequences of the verification of completeness in clinical practice guideline modeling : a theoretical and empirical study with hypertension. *AMIA Symposium*, **2009**, 60–4.

BURGES C. J. C., SHAKED T., RENSHAW E., LAZIER A., DEEDS M., HAMILTON N. & HULLENDER G. N. (2005). Learning to rank using gradient descent. In *ICML*, volume 119 of *ACM International Conference Proceeding Series*, p. 89–96 : ACM.

DEKEL O., MANNING C. D. & SINGER Y. (2003). Log-linear models for label ranking. In *NIPS*, p. 497–504 : MIT Press.

FÜRNKRANZ J. & HÜLLERMEIER E. (2010). *Preference learning : An introduction*.

HAN Q., MARTINEZ DE RITUERTO DE TROYA I., JI M., GAUR M. & ZEJNILOVIC L. (2018). A collaborative filtering recommender system in primary care : Towards a trusting patient-doctor relationship. In *IEEE International Conference on Healthcare Informatics (ICHI)*, p. 377–379.

KORICHE F. & ZANUTTINI B. (2010). Learning conditional preference networks. *Artificial intelligence*, **174**(11), 685–703.

LAMY JB (2016). Ontology-Oriented Programming for Biomedical Informatics. *Studies in health technology and informatics (STC)*, **221**, 64–68.

LAMY JB (2017). Owlready : Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. *Artif Intell Med*, **80**, 11–28.

LAMY JB (2019). *Advances in nature-inspired computing and applications*, chapter Artificial Feeding Birds (AFB) : a new metaheuristic inspired by the behavior of pigeons, p. 43–60. Springer.

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

NGOC K. A. P., LEE Y. K. & LEE S. Y. (2005). OWL-based user preference and behavior routine ontology for ubiquitous system.

SEZGIN E. & OZKAN S. (2013). A systematic literature review on health recommender systems. In *E-Health and Bioengineering Conference (EHB)*, p. 1–4.

SMITH MK (2003). Michael Polanyi and tacit knowledge. *The encyclopedia of informal education*.

SUBRAMANIYASWAMY V., VARADHARAJAN V. & INDRAGANDHI V. (2013). A review of ontology-based tag recommendation approaches. *International Journal of Intelligent Systems*, **28**(11), 1054–1071.

TEHRANI A. F., CHENG W. & HULLERMEIER E. (2012). Preference learning using the Choquet integral : The case of multipartite ranking.

TSAI K. H., CHIU T. K., LEE M. C. & WANG T. I. (2006). A learning objects recommendation model based on the preference and ontological approaches.

TSOPRA R., LAMY J. B. & SEDKI K. (2018). Using preference learning for detecting inconsistencies in clinical practice guidelines : methods and application to antibiotherapy. *Artif Intell Med*, **89**, 24–33.

TSOPRA R., SEDKI K., COURTINE M., FALCOFF H., DE BÉCO A., MADAR R., MECHAÏ F. & LAMY J. B. (2019). Helping GPs to extrapolate guideline recommendations to patients for whom there are no explicit recommendations, through the visualization of drug properties. The example of AntibioHelp® in bacterial diseases. *J Am Med Inform Assoc*, **accepted**.

TSOPRA R., VENOT A. & DUCLOS C. (2014a). An algorithm using twelve properties of antibiotics to find the recommended antibiotics, as in CPGs . In *AMIA Annu Symp Proc*, volume 1115-24.

TSOPRA R., VENOT A. & DUCLOS C. (2014b). Towards evidence-based CDSSs implementing the medical reasoning contained in CPGs : application to antibiotic prescription. In *Stud Health Technol Inform*, volume 205, p. 13–7.

WANG T. I., TSAI K. H., LEE M. C. & CHIU T. K. (2007). Personalized learning objects recommendation based on the semantic-aware discovery and the learner preference pattern. *Educational technology & society*, **10**(3), 84–105.

WERNER D., SILVA N., CRUZ C. & BERTAUX A. (2014). An ontology-based recommender system using hierarchical multiclassification for economical e-news. In *Proceedings of the International Conference on Informatics in Economy (IE 2014)*, Bucharest, Romania.

YANG XS (2010). *Nature-inspired metaheuristic algorithms (second edition)*. Luniver Press.

Explorer la synergie entre le Web sémantique et la vision par ordinateur pour la personnalisation

Vincent Lully[1], feu Philippe Laublet[1], Milan Stankovic[1,2], Filip Radulovic[2]

[1] Sorbonne Université, 28, rue Serpente, 75006 Paris, France
[2] Sépage, 38, avenue de l'Opéra, 75002 Paris, France

Résumé : Nous explorons la synergie entre le Web sémantique et la vision par ordinateur dans le contexte des systèmes de personnalisation. Nous présentons deux nouvelles applications : profilage utilisateur à partir des images et sélection d'images dans les bannières de recommandation.

Mots-clés : Web sémantique, Vision par ordinateur, Personnalisation, Profil utilisateur, Image.

Plusieurs travaux ont montré des points de convergence intéressants entre le domaine du Web sémantique et celui de la vision par ordinateur : amélioration de la détection des objets à l'aide des graphes de connaissances externes (Fang *et al.*, 2017), description des scènes avec des triplets (Baier *et al.*, 2017), complétion des graphes de connaissances avec des caractéristiques visuelles (Thoma *et al.*, 2017) et recherche visuo-sémantique (Ferrada *et al.*, 2017). Nous explorons la synergie entre les deux domaines dans le contexte des systèmes de personnalisation. Nous présentons deux nouvelles applications.

La première application est le profilage utilisateur à partir des images. Nous avons proposé deux approches. La première approche consiste à associer une image aux entités sémantiques correspondant aux objets apparaissant dans l'image. La seconde approche consiste à associer une image aux entités sémantiques dépeintes par les images visuellement similaires et qui existent dans la portée conceptuelle du jeu de données au sein duquel la personnalisation est réalisée. Notre évaluation a montré la supériorité de notre seconde approche par rapport au référentiel en termes de la pertinence des entités constituant les profils.

La deuxième application est la sélection d'images dans les bannières de recommandation. Nous ordonnons les images associées au produit recommandé et affichons dans les bannières celle correspondant le mieux au profil de l'utilisateur dans le but d'augmenter son affinité envers le produit. Nous avons mené une étude utilisateur avec 32 participants en utilisant un vrai jeu de données contenant 1,357 circuits touristiques dans 136 pays accompagnés de 11,614 images distinctes. Les résultats ont montré la performance prometteuse de notre approche en termes de persuasion, attention, efficacité et affinité.

Les travaux résumés ci-dessus ont été originellement publiés dans un article au congrès SEMANTICS 2018 (Lully *et al.*, 2018).

Références

BAIER S., MA Y. & TRESP V. (2017). Improving visual relationship detection using semantic modeling of scene descriptions. In *International Semantic Web Conference*, p. 53–68 : Springer.

FANG Y., KUAN K., LIN J., TAN C. & CHANDRASEKHAR V. (2017). Object detection meets knowledge graphs. In *Proceedings of IJCAI'17*, p. 1661–1667 : AAAI Press.

FERRADA S., BUSTOS B. & HOGAN A. (2017). Imgpedia : a linked dataset with content-based analysis of wikimedia images. In *International Semantic Web Conference*, p. 84–93 : Springer.

LULLY V., LAUBLET P., STANKOVIC M. & RADULOVIC F. (2018). Exploring the synergy between knowledge graph and computer vision for personalisation systems. In *Proceedings of the 14th International Conference on Semantic Systems*, volume 137, p. 175–186 : Elsevier.

THOMA S., RETTINGER A. & BOTH F. (2017). Towards holistic concept representations : embedding relational knowledge, visual attributes, and distributional word semantics. In *International Semantic Web Conference*, p. 694–710 : Springer.

SaGe : préemption Web pour les services publics d'évaluation de requêtes SPARQL⋆

Thomas Minier, Hala Skaf-Molli and Pascal Molli

LS2N – University of Nantes, Nantes, France
{thomas.minier,hala.skaf,pascal.molli}@univ-nantes.fr

Suivant les principes du Linked Open Data (LOD), les fournisseurs de données hébergent publiquement des millions de triples au format RDF. Cependant, fournir un service public qui permet à n'importe qui d'exécuter n'importe quelle requête SPARQL sur ces données est toujours un problème ouvert. Comme ces services sont soumis à une charge imprévisible de requêtes, le défi est d'assurer qu'ils demeurent *stables* malgré des variations en termes de taux d'arrivées des requêtes et des ressources nécessaires à leur évaluation.

Pour résoudre ce problème, la plupart des fournisseurs de données appliquent une politique d'utilisation équitable des serveurs basée sur des *quotas*. Ces derniers visent à empêcher les *effets convois* (Blasgen *et al.*, 1979), *c.a.d*, une requête longue à exécuter bloque l'évaluation des autres. Le principal défaut de cette politique est qu'elle empêche les requêtes interrompues de délivrer des résultats complets. Cela constitue une limite sérieuse pour les utilisateurs du LOD, qui peuvent vouloir exécuter des requêtes longues.

Nous pensons que le problème lié aux quotas ne réside pas dans l'interruption des requêtes, mais dans l'impossibilité pour les clients de *reprendre leur exécution* après interruption. Néanmoins, il n'existe pas de modèle de préemption pour le Web qui permet la suspension et la reprise de l'exécution de requêtes SPARQL. Dans (Minier *et al.*, 2019), nous proposons SaGe, un moteur d'évaluation de requête SPARQL basé sur la *préemption Web*. Il permet à un serveur Web de suspendre une requête en cours d'exécution après un certain temps, puis de reprendre son exécution ultérieurement. Une fois suspendue, l'état d'une requête est retourné au client, qui peut reprendre son exécution en renvoyant l'état au serveur.

La préemption Web engendre des coûts supplémentaires pour le serveur Web, qui doit suspendre la requête courante puis reprendre l'exécution de la suivante. En conséquences, le problème scientifique majeur est de maintenir ce surcoût marginal, quelle que soit la requête, afin d'assurer une exécution performante. Nos contributions sont les suivantes :

— Nous formalisons le modèle de *préemption Web* qui permet de suspendre et de reprendre l'exécution de requêtes SPARQL. Nous définissons aussi un ensemble d'opérateurs d'exécution préemptifs, dont la complexité d'arrêt et de reprise est bornée.

— Nous proposons SaGe, un moteur d'évaluation de requêtes SPARQL composé d'un serveur Web préemptif et d'un client Web intelligent qui permet l'évaluation de n'importe quelle requête SPARQL. Nous comparons ensuite les performances de ce nouveau moteur de requêtes avec les approches existantes. Nos résultats expérimentaux démontrent que SaGe surpasse de plusieurs ordres de grandeurs les approches existantes en termes de temps moyen d'exécution des requêtes et de temps d'obtention des premiers résultats.

Références

Blasgen M. W., Gray J., Mitoma M. F. & Price T. G. (1979). The convoy phenomenon. *Operating Systems Review*, **13**(2), 20–25.

Minier T., Skaf-Molli H. & Molli P. (2019). SaGe : Web Preemption for Public SPARQL Query Services. In *The World Wide Web Conference 2019 (WWW'19)*, San Francisco, United States.

⋆. Article complet publié sous le titre "SaGe: Web Preemption for Public SPARQL query services" dans les actes de *The World Wide Web Conference 2019 (WWW'19)*.

LDP-DL : Une langage pour definir la conception des Linked Data Platforms

Noorani Bakerally[1], Antoine Zimmermann[2], Olivier Boissier[2]

[1] CNRS, LAAS, 7 avenue du Colonel Roche,F-31400 Toulouse, France
noorani.bakerally@gmail.com

[2] Univ Lyon, IMT MinesSaint- Etienne, CNRS, Laboratoire Hubert Curien UMR 5516, F-42023 Saint- Etienne, France
antoine.zimmermann@emse.fr,olivier.boissier@emse.fr

Linked Data Platform 1.0 (LDP) [2], recommandation du W3C, spécifie comment exposer des données liées en conformité avec les principes REST. Son utilisation améliore l'interopérabilité, facilite l'exploitation des données, notamment dans des écosystèmes d'information comme l'*open data* ou pour l'échange d'information inter-organisation à l'échelle du Web. Tandis que LDP (et les technologies du Web sémantique) peut grandement simplifier le travail des utilisateurs de la donnée, cela impose un lourd fardeau aux éditeurs de données, qui doivent utiliser des technologies souvent méconnues. Le passage aux LDP nécessite une refonte de la manière dont les données sont organisées et publiées. En outre, les implémentations actuelles en sont à leurs débuts et ne fournissent pas de support automatique pour le déploiement de données, qu'elles soient statiques, dynamiques ou hétérogènes. À cette fin, nous fournissons une approche dont le noyau est un langage [1], LDP-DL, pour spécifier comment les sources de données existantes doivent être utilisées pour générer des LDP de manière indépendante et compatible avec toute implémentation du standard, et déployable sur chacune d'elles. En bref, un document en LDP-DL, appelé document de conception, fournit une description abstraite des *ressources LDP*, qui peuvent être des *sources RDF* ou bien des *sources non-RDF*. Celles-ci peuvent être organisées dans des *conteneurs LDP*, eux-mêmes des sources RDF, pouvant contenir des ressources LDP appelées ses membres. Des trois types de conteneurs existant, nous ne considérons que les *conteneurs basiques* qui ne contiennent que des documents Web. LDP-DL fournit les constructeurs `ContainerMap` et `NonContainerMap` qui permettent de décrire, en fonction des données sources, quelles seront les ressources (conteneurs ou non) qui seront accessibles via la plateforme. Pour chaque ressource, le constructeur `ResourceMap` sert à définir quel graphe RDF lui est associé. Enfin, un `ContainerMap` peut faire référence à d'autres `ContainerMaps` et `NonContainerMaps` afin de constituer une hiérarchie de conteneurs imbriqués. Dans le cadre de notre approche, nous fournissons également un *workflow* pour la génération de LDP comprenant deux processus principaux : la *LDP-isation* et le *déploiement*. Durant le processus de LDP-isation, LDP-DL est utilisé pour décrire la conception de la plateforme en terme de structuration et organisation des données. Le document de conception est traité vis-à-vis des sources de données existantes par un interpréteur LDP-DL, ou *LDP-isateur*. Le résultat est un *jeu de données LDP* contenant les ressources LDP et leurs graphes associés. Ensuite, le processus de déploiement met à disposition les données sur une plateforme concrète, soit en ajoutant les données, ressource par ressource, à une implémentation existante de LDP via des requêtes HTTP POST ; soit à l'aide de notre implémentation qui peut consommer directement un jeu de données LDP. Dans ce dernier cas, nous avons prévu un mode de LDP-isation et de déploiement intermédiaire qui permet de récupérer, au moment de la requête utilisateur, une ressource LDP générée à la volée en fonction des données sources. Ce mode de fonctionnement permet de toujours fournir un contenu informationnel à jour par rapport aux données sources initiales. L'ensemble de la chaine de traitement a été implémentée et testée sur des données statiques, dynamiques et hétérogènes.

Références

[1] BAKERALLY N., ZIMMERMANN A. & BOISSIER O. (2018). LDP-DL : A language to define the design of linked data platforms. In *European Semantic Web Conference*, p. 33–49 : Springer.

[2] SPEICHER S., ARWE J. & MALHOTRA A. (2015). *Linked Data Platform 1.0*. Rapport interne, World Wide Web Consortium (W3C).

Booster le matching holistique : un jeu d'alignement de référence basé sur les cliques relaxés

Philippe Roussille, Imen Megdiche, Olivier Teste, Cassia Trojahn

IRIT UMR 5505, Institut de Recherche en Informatique de Toulouse, Toulouse, France
firstname.lastname@irit.fr

Résumé

L'alignement de deux ressources ontologiques ou *pairwise matching* est une thématique largement étudiée par la communauté scientifique. Malgré les avancées dans ce domaine, la généralisation de ces approches pour prendre en compte plusieurs sources simultanément ou ce que l'on appelle holistic matching, demeure un domaine de recherche et d'expérimentation plus que important dans l'ère des *Data Lakes* et du *Big Data*. Depuis quelques années, on commence à voir émerger de nouvelles approches et outils, pour le matching holistique, sans pour autant être en capacité de valider expérimentalement l'efficacité de ces systèmes les uns par rapport aux autres. La mise à disposition d'alignements de références pour le matching holistique a fait l'objet de nos travaux (Roussille *et al.* (2018)). Nous avons présenté une méthodologie pour construire un jeu d'alignements de référence *pseudo*-holistique à partir de jeux d'alignements de références existants de type pairwise. Nos alignements se présentent sous forme de cliques de graphes impliquant un nombre variable d'ontologies. Ces alignements appartiennent à deux niveaux de relaxation, permettant de capter la complexité sémantique qui peut introduire des contradictions logiques dans les structures complètes.

— Le premier niveau est qualifié de 'strict'. L'idée étant de chercher des sous-graphes complets à partir de N graphes en entrée.
— Le deuxième niveau est 'relaxé', nous avons proposé deux méthodes :
  — La première méthode est une relaxation directe des cliques à N noeuds. Nos alignements sont des sous-graphes incomplets.
  — La deuxième méthode se base sur les relations intra-ontologies. Le nombre de graphes en entrée $\in [0, N]$

Nous avons évalué notre outil, Holontology, conçu pour générer des alignements holistiques, sur le jeu de données de la conférence OAEI 2018 [1]. Holontology obtient de meilleurs résultats que les approches dites *pair à pair* lorsque le nombre d'ontologies alignées est supérieur à deux. Bien que nous proposons une approche pseudo-holistique, étant donné que nous partons d'alignements de référence de type pairwise, il s'agit d'un premier pas vers l'évaluation des systèmes holistiques. Pour les travaux futurs, nous prévoyons d'étendre l'évaluation de cet ensemble de données à une vérification manuelle par des experts, ainsi que de travailler sur la fermeture transitive des alignements obtenus, mais aussi de généraliser les types de relations utilisées autres que les relations d'équivalence.

Références

ROUSSILLE P., MEGDICHE I., TESTE O. & TROJAHN C. (2018). Boosting holistic ontology matching : Generating graph clique-based relaxed reference alignments for holistic evaluation. In *Knowledge Engineering and Knowledge Management - 21st International Conference, EKAW 2018, Nancy, France, November 12-16, 2018, Proceedings*, p. 355–369.

1. le track Conference de Ontology Alignment Evaluation Campaign (OAEI)

Approche de prédiction de présence d'amiante dans les bâtiments basée sur l'exploitation des descriptions temporelles incomplètes de produits commercialisés

Thamer Mecharnia[1,2], Nathalie Pernelle[1], Lydia Chibout Khelifa[2], Fayçal Hamdi[3]

[1] LRI, Université Paris sud, Orsay, France
prenom.nom@lri.fr

[2] CENTRE SCIENTIFIQUE ET TECHNIQUE DU BÂTIMENT (CSTB), Champs sur Marne, France
prenom.nom@cstb.fr

[3] CEDRIC – CNAM, Paris, France
faycal.hamdi@cnam.fr

Résumé : La production, l'importation et la commercialisation d'amiante sont interdites depuis le premier janvier 1997 en France. Cependant, il en reste des millions de tonnes disséminés dans les usines, les immeubles, les établissements scolaires, ou encore les hôpitaux.

Dans cet article nous proposons une méthode de prédiction de présence d'amiante basée sur des données temporelles décrivant la probabilité de présence d'amiante dans des produits commercialisés pour calculer une probabilité d'existence de produits amiantés dans les bâtiments. Pour atteindre notre but, nous proposons une ontologie amiante qui va être peuplée en utilisant les données issues de ressources externes. Ensuite, ces informations sont utilisées pour calculer la probabilité d'amiante pour les éléments constituant un bâtiment donné. Notre approche a été expérimentée sur des données synthétiques décrivant 120 bâtiments en s'appuyant sur les 704 produits amiantés de l'INRS et l'ANDEVA.

Mots-clés : Ontologies, Informations incertaines, données temporelles, prédiction.

1 Introduction

Pour ses qualités ignifuges, la France a abondamment utilisé l'amiante, de l'avant-guerre à son interdiction en 1997, plutôt tardive en Europe, en particulier au cours des décennies 1950 à 1970. La nocivité de l'amiante ou de l'asbeste est connue de longue date, mais son danger est identifié depuis le début du XXe siècle. Une accumulation de données scientifiques et médicales sur l'amiante telle que rapportée dans l'expertise collective de l'INSERM de 1997, rappelle les principales étapes des connaissances médicales relatives aux effets des expositions à l'amiante sur la santé.

La production, l'importation et la commercialisation d'amiante sont interdites depuis le premier janvier 1997 en France. Cependant, il en reste des millions de tonnes disséminés dans les usines, les immeubles, les établissements scolaires, ou encore les hôpitaux. Le repérage de parties amiantées est donc d'importance que ce soit pour réaliser des travaux de mise en conformité ou pour envisager le recyclage des éléments du bâtiment (e.g. fenêtre, plancher, porte, ...) dans le cadre de l'économie circulaire. Dans le cadre des travaux du Plan de recherche et développement Amiante (PRDA), le CSTB a été sollicité pour élaborer un outil en ligne fournissant une aide au repérage de matériaux amiantés dans les bâtiments (ORIGAMI) qui a pour objectif d'orienter l'opérateur de repérage et de l'aider dans la préparation de son programme de repérage. Il ne se substituera en aucun cas au repérage des matériaux et produits contenant de l'amiante réalisé par un professionnel conformément à la norme NF X 46-020. L'outil pourra également être un support de formation, voire être étendu à un usage « Particuliers » dans un objectif de sensibilisation au risque amiante.

Ce projet a fait émerger de nouvelles problématiques à savoir, la pérennisation des connaissances dans le domaine de l'amiante, le partage et la réutilisation de ces connaissances

dans d'autres domaines tels que le réemploi et le recyclage des produits/matériaux pour la construction (comme le béton dans le cadre de l'économie circulaire). Répondre à ces problématiques devrait apporter une aide dans le domaine de l'emploi ou de la restriction des matériaux/produits (à valider par les Groupes de Travail Spécialisés du CSTB) ainsi que la détection de cas particuliers.
Dans cet article nous proposons une méthode d'analyse de données temporelles incertaines qui permet de calculer une probabilité d'existence de produits amiantés dans les bâtiments à partir d'un ensemble de descriptions de bâtiments et de ressources externes décrivant des produits ayant été amiantés ou probablement amiantés à certaines périodes. Pour atteindre notre but, nous proposons une ontologie amiante qui est enrichie et peuplée en utilisant les données d'entrée. Ensuite, cette ontologie peuplée est utilisée pour le calcul de la probabilité de présence d'amiante dans un élément constitutif d'un bâtiment. Notre approche a été expérimentée sur des données synthétiques décrivant 120 bâtiments en s'appuyant sur les 704 produits amiantés décrits par l'INRS et l'ANDEVA.

Dans la section 2, nous présentons les ressources disponibles au sein du CSTB et la problématique de prédiction que nous avons définie en collaboration avec les experts. Dans la section 3, nous décrivons l'ontologie amiante proposée. Puis, en section 4, nous présentons l'approche que nous avons définie pour enrichir et peupler l'ontologie par les produits commercialisés décrits dans l'INRS et l'ANDEVA et pour calculer la probabilité de l'existence de l'amiante dans une partie du bâtiment. Finalement, en section 5, nous présentons les résultats obtenus dans nos premières expérimentations.

2 Contexte et Problématique

Dans le problème que nous traitons, l'objectif est de prédire la présence potentielle d'amiante dans un bâtiment. Le CSTB archive un ensemble de documents qui décrivent les bâtiments construits en France, dont les deux types de documents décrits ci-dessous :

— **Projet type homologué** : document qui contient un ensemble d'informations décrivant un bâtiment (nom, adresse, type, année de construction, région), et la liste de ses structures (ouverture extérieure, balcon, ...). Pour chaque structure, le document décrit l'ensemble de ses localisations (porte, fenêtre, ...) ainsi que les familles de produits utilisées pour chaque localisation (enduit, colle, ...).

— **Diagnostic amiante** : document qui décrit les résultats des prélèvements effectués sur un bâtiment pour détecter la présence d'amiante dans des éléments constituant des parties de bâtiments (produits, localisations).

A l'heure actuelle, pour prédire une présence éventuelle d'amiante et demander un prélèvement, l'expert utilise les diagnostics effectués pour d'autres projets homologués. Si le bâtiment possède les mêmes caractéristiques que celui qui est mentionné dans un diagnostic (même région et même type), et si une classe de produit contient de l'amiante, il suppose que la même classe de produit peut contenir de l'amiante et il demande le prélèvement et l'analyse d'un échantillon. Dans les cas où il ne trouve pas de diagnostic qui concerne un bâtiment similaire au bâtiment en cours d'étude, il demande que des échantillons soient analysés en laboratoire pour toutes les parties du bâtiment.

L'objectif du projet est d'aider l'expert à prédire la présence de l'amiante dans les familles de produits dans un bâtiment donné en s'appuyant sur les ressources documentaires du CSTB qui couvre la période de 1943 à 1997. Comme les projets types homologués ne mentionnent pas les produits commercialisés réellement utilisés lors de la construction du bâtiment, nous faisons l'hypothèse que nous pouvons calculer une probabilité d'existence d'amiante pour un produit utilisé à partir des produits de cette famille commercialisés au moment de la construction de ce bâtiment, et que les familles de produits n'utilisent plus d'amiante conformément à l'interdiction de son utilisation à partir de 1997. Plus précisément, les différentes étapes du projet sont les suivantes :

(1) Construire une ontologie Amiante qui permette de modéliser les connaissances sur les bâtiments et les diagnostics réalisés sur les bâtiments quand ils existent.

(2) Enrichir et peupler l'ontologie par des classes et des instances de classes et de propriétés issues d'un processus d'extraction automatique des informations décrites dans les projets homologués.

(3) Enrichir l'ontologie en s'appuyant sur des ressources externes décrivant les produits commercialisés et la probabilité de présence d'amiante dans ces produits.

(4) Proposer une approche de prédiction de présence d'amiante basée sur ces connaissances.

Nous disposons pour l'instant d'un nombre insuffisant de projets homologués et de diagnostics pour définir une méthode générique pour l'étape (2) d'extraction automatique, ni pour définir une méthode supervisée qui permette d'apprendre à prédire la présence d'amiante à partir des bâtiments, des diagnostics associés et des produits commercialisés pour l'étape (4). Aussi, dans cet article, nous nous focalisons sur les étapes (1), (3) et sur la définition d'une approche non supervisée pour l'étape (4).

3 L'Ontologie Amiante

Nous avons manuellement construit la partie haute de l'ontologie Amiante en nous basant sur les ressources documentaires du CSTB, les besoins en termes de prédiction (cf. figure 1) et en interagissant avec l'expert. Les principaux concepts de cette ontologie sont les suivants :

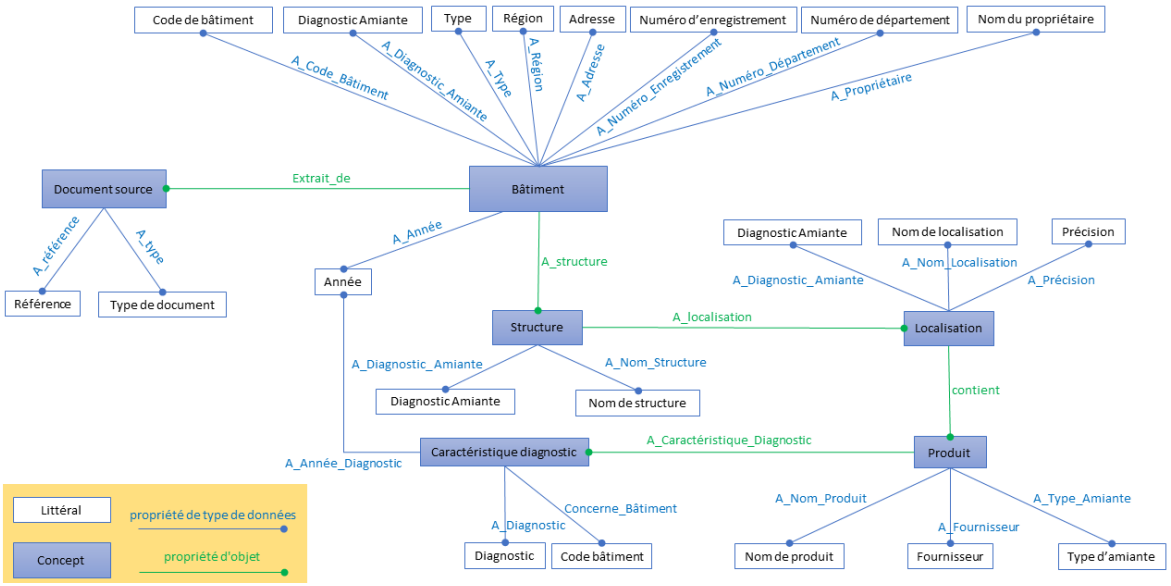


FIGURE *1 – Concepts principaux de l'ontologie Amiante*

- Bâtiment : construction qui est caractérisée par un code, un type (codification CSTB qui correspond à un genre donné de bâtiment : école, logement), une année de construction, une adresse, et la région où le bâtiment a été construit.
- Structure : composant essentiel d'un bâtiment qui correspond à un sous-espace du bâtiment (ex : terrassement, balcon, escalier, toiture, plancher-sol, ...)
- Localisation : désigne un élément de base qui appartient à une structure de bâtiment (ex : porte, fenêtre, plancher, mur, ...).
- Produit : décrit un produit qui peut entrer dans la composition des localisations (ex : colle, enduit, . . .). Un produit est décrit par son nom, le nom de fournisseur et le type d'amiante s'il est amianté.

— Document source : référence des documents, projet type homologué ou diagnostic Amiante, qui décrivent respectivement les propriétés des Bâtiments, les familles de produits ou les résultats des prélèvements effectués sur le bâtiment. Un document source est caractérisé par son type (projet type homologué ou diagnostic amiante) et sa référence au sein du CSTB.
— Caractéristique diagnostic : est composée des informations extraites à partir des diagnostics. Elle contient le résultat de l'existence d'amiante dans un bâtiment similaire dans la même année, le code de ce bâtiment et l'année de sa construction.

Pour représenter les résultats des diagnostics effectués sur les produits, quand ils existent, nous avons modélisé le concept de Caractéristique diagnostic qui est décrit par le résultat de l'existence d'amiante propriété $A\_Diagnostic$ qui prend la valeur *oui* ou *non*, pour un code de bâtiment et une année de construction donnée. Les autres éléments de bâtiment (i.e. localisation, structure) et le bâtiment lui-même sont également décrits par une propriété $A\_Diagnostic\_Amiante$ qui prend la valeur *oui* ou *non* selon que l'un de ses éléments contient de l'amiante.

4 Approche prédictive

Dans les documents du CSTB, nous ne connaissons que les classes de produit utilisées pour une localisation mais nous ne connaissons pas la référence du produit utilisé. Par exemple, nous savons que le produit utilisé est un « Enduit » mais nous ne savons pas de quel enduit (i.e. produit commercialisé) il s'agit exactement.

Dans ce qui suit, nous proposons d'utiliser une méthode non-supervisée qui se base sur la présence de l'amiante dans les produits commercialisés pour prédire la probabilité de l'existence d'amiante pour les produits d'un bâtiment construit à une date donnée, puis plus globalement pour ses localisations, ses structures et le bâtiment lui-même. Dans une première étape, nous utilisons les ressources externes de l'ANDEVA (Association Nationale de Défense des Victimes de l'Amiante) et de l'INRS (Institut National de Recherche et de Sécurité) pour enrichir l'ontologie Amiante par des produits commercialisés, puis nous utilisons ces informations pour calculer la probabilité de présence d'amiante en utilisant un modèle de graphe probabiliste.

4.1 Enrichissement et peuplement de l'ontologie des produits commercialisés

Nous utilisons les deux ressources externes fournies par l'ANDEVA et l'INRS pour enrichir l'ontologie avec des sous-classes de produit et des instances de produits commercialisés avec leurs caractéristiques amiante. L'ANDEVA publie sur son site web [1] une liste des matériaux amiantés. Cette liste est représentée sous forme d'un tableau qui contient : la classe du produit, un nom ou un ensemble de noms de produits qui partagent les mêmes propriétés avec leur nom de fournisseur, et l'année à partir de laquelle ils ne sont plus amiantés. L'INRS publie également une liste [2] des matériaux amiantés représentés sous forme d'un tableau qui contient un ensemble de noms de produits, le nom éventuel du fournisseur, les intervalles de temps où le produit est amianté avec un degré d'incertitude, le type d'amiante, et les types d'utilisation. Plus précisément, dans la ressource INRS, les différents intervalles de temps peuvent correspondre à différentes annotations décrivant la présence d'amiante : le produit est amianté, la présence d'amiante est inconnue (non renseignée), le produit n'est plus amianté, le produit n'est plus commercialisé.

Extraction automatique des classes de produits et des descriptions de produits dans les données tabulaires

Dans un premier temps, nous extrayons automatiquement les informations concernant les produits de l'ANDEVA et l'INRS en nous basant sur la structuration des données et sur

1. http://andeva.fr/?-Liste-de-produits-contenant-de-l-
2. http://www.inrs.fr/media.html?refINRS=ED%201475

des expressions régulières pour l'extraction des intervalles et des probabilités de présence d'amiante dans les annotations. Puis, nous enrichissons l'ontologie avec les classes des produits qui deviennent des sous-classes de la classe $Produit$ de l'ontologie Amiante. Pour représenter les caractéristiques produits issues de chacune des sources, ainsi que les caractéristiques issues de la fusion de ces informations, nous utilisons la réification pour représenter les informations temporelles et les probabilités et ajoutons à l'ontologie Amiante représentée en figure 1 un nouveau concept de $Caracteristique\ Extraite$ qui va permettre de représenter les caractéristiques de présence d'amiante dans un matériau, en précisant l'intervalle de temps, la probabilité de la présence d'amiante, et la source de cette caractéristique (INRS, ANDEVA ou fusion) (cf. figure 2). Nous indiquons une probabilité de 1 pour les produits amiantés sur la période, de 0.5 quand la présence d'amiante est inconnu et de 0 sinon.

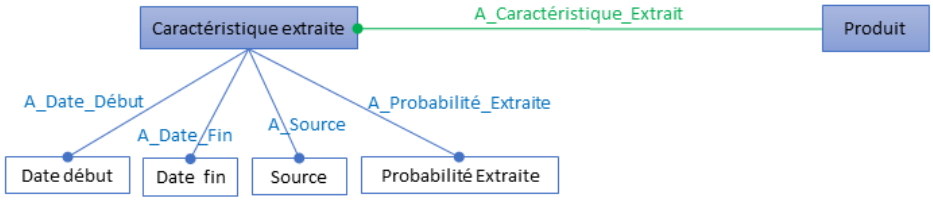


FIGURE 2 – *Représentation des caractéristiques extraites dans l'ontologie Amiante*

Prenons l'exemple de l'« ARMAZOL » qui est décrit dans l'INRS et l'ANDEVA par les tableaux 1 et 2 respectivement, et qui ne comporte pas de mention de fournisseur.

TABLE *1 – Extrait de l'INRS : « ARMAZOL »*

Produit	Fournisseur	Renseignements divers	Type d'amiante	Type d'utilisation
ARMAZOL		Amianté jusqu'en 1982, non renseigné après	Amiante	Revêtement de sols en dalles ou en rouleaux

TABLE *2 – Extrait de l'ANDEVA : « ARMAZOL »*

Nom de famille de produits	Produit et fournisseur	Amianté jusqu'en
Revêtements de sols en dalles ou en rouleaux	ARMAZOL	1990

A l'issue de cette étape, la classe de produit « Revêtements de sols en dalles ou en rouleaux » est créée et deux instances de cette classe dont le nom est « ARMAZOL » sont créées qui ont les caractéristiques suivantes :

— Le premier produit est présenté comme provenant de la source INRS, et comme étant amianté (probabilité = 1) de 1946 à 1982, puis comme potentiellement amianté jusqu'en 1997 (probabilité = 0,5).
— Le second produit est présenté comme provenant de la source ANDEVA, il est amianté de 1946 à 1990 (probabilité = 1), puis non amianté (probabilité = 0) jusqu'en 1997.

Fusion des descriptions de produits identiques

Dans un deuxième temps, nous fusionnons les descriptions de produits identiques. Pour cela, nous devons décider que ces descriptions réfèrent bien au même produit (i.e. liage de données) puis résoudre les éventuels conflits quand les sources de données ne s'accordent pas

sur certaines valeurs de propriété.
L'étape de liage de données s'effectue simplement en se basant sur le nom commercial du produit (i.e. sur l'égalité de chaines de caractères qui ne diffèrent pas selon les deux sources).

Lors de la fusion des deux descriptions, nous fusionnons automatiquement les valeurs des propriétés d'un produit en faisant l'union des valeurs distinctes. Cependant, dans le cas des intervalles correspondant aux caractéristiques amiante, les intervalles et les degrés de présence d'amiante peuvent être différents dans l'INRS et l'ANDEVA, comme c'est le cas dans l'exemple de l'Armazol décrit précédemment. Aussi, nous effectuons la fusion de la manière suivante :

Nous faisons l'union ordonnée des bornes des intervalles de temps successifs de l'ANDEVA et l'INRS, en éliminant les doublons, et pour chaque paire successive de bornes, nous créons un intervalle de temps et associons à cet intervalle une probabilité de présence d'amiante qui correspond au degré de présence d'amiante le plus élevé des deux ressources. Les deux ressources étant de même fiabilité, nous appliquons ici une approche pessimiste qui considère le plus haut degré de présence d'amiante et c'est ce calcul qui sera utilisé dans la prédiction de présence d'amiante dans le bâtiment.

Ainsi, la fusion des caractéristiques amiante du produit Armazol (schématisé en figure 3) conduit aux étapes suivantes :

1. Après avoir ordonné les bornes des intervalles d'entrée des deux sources, on obtient : {1946, 1982, 1990, 1997}.
2. A partir de ces bornes, nous construisons les intervalles du résultat : {[1946, 1982[, [1982, 1990[, [1990, 1997[}.
3. Les intervalles [1982, 1990[et [1990, 1997[contiennent des informations contradictoires : probabilité = 0.5 pour l'INRS et probabilité = 1 pour l'ANDEVA pour le premier intervalle et probabilité = 0,5 et probabilité = 0 pour le deuxième intervalle. Pour résoudre ce conflit, nous prenons le maximum des deux valeurs (probabilité = 1 pour le premier intervalle et 0,5 pour le deuxième).

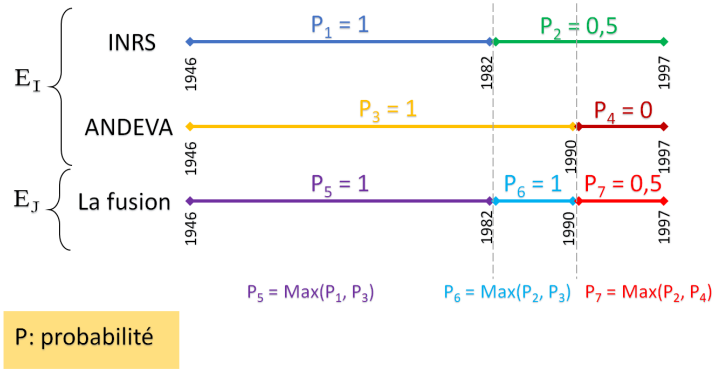


FIGURE *3 – Exemple de fusion des intervalles de temps et des probabilités de présence d'amiante*

4.2 Calcul de la probabilité de présence d'amiante pour un produit utilisé dans un bâtiment

Comme les documents ne mentionnent pas les produits réellement utilisés lors de la construction d'un bâtiment, nous faisons l'hypothèse que nous pouvons calculer la probabilité d'existence d'amiante pour un produit utilisé à partir des produits de la même classe commercialisés au moment de la construction de ce bâtiment, en considérant qu'ils sont équiprobables.

TABLE *3 – Caractéristiques des produits appartenant à la classe "Revêtements muraux"*

Produit	Amianté jusqu'en
COLOVINYL	1983
DECOVER	/
NOVILON	1981
STRAIRLAM	/

TABLE *4 – Probabilités des produits de la classe "Revêtements muraux" en 1982*

Produit	Probabilité
COLOVINYL	1
DECOVER	0,5
NOVILON	0
STRAIRLAM	0,5

Ainsi, pour chaque produit p_k appartenant à la classe $F(p_k)$ qui est utilisé dans un bâtiment construit à une date d, nous calculons une probabilité de présence d'amiante $p_a(p_k, d)$ en sommant les probabilités fusionnées p_{ext} des produits commercialisés p_j de type $F(p_k)$ qui sont amiantés à cette date et divisons cette somme sur le nombre total de produits de cette famille qui étaient en cours d'utilisation à cette date :

$$p_a(p_k, d) = \frac{\sum_{p_j \in F(p_k)} p_{ext}(p_j, d)}{\mid p_j \mid} \tag{1}$$

où :
$p_a(p_k, d)$ est la probabilité amiante calculée pour le produit inconnu p_k à la date d ,
$p_{ext}(p_j, d)$ est la probabilité amiante après le processus de fusion du produit commercialisé p_j et qui est de même type F que p_k.

Par exemple : pour calculer la probabilité de l'existence d'amiante dans un produit de la classe de produits "Revêtements muraux" en 1982, nous allons utiliser les informations des ressources externes (ANDEVA et INRS) pour appliquer notre formule 1. La classe de produits "Revêtements muraux" contient les quatre produits décrits dans le tableau 3. Maintenant, nous remplaçons les variables dans l'équation 1, tel que $d = 1982$, $p_k = RevêtementsMuraux$ et le nombre de produits $\mid p_j \mid = 4$:

$$p_a(RevêtementsMuraux, 1982) = \frac{\sum_{p_j \in F(RevêtementsMuraux)} p_{ext}(p_j, 1982)}{4}$$

En 1983, nous disposons des probabilités de produits montrées dans le tableau 4 :
— "COLOVINYL" qu'est toujours amianté avant 1983, a la probabilité 1.
— Dans le cas de "DECOVER" et "STRAIRLAM" où nous n'avons pas des informations sur la présence d'amiante, et donc nous mettons la probabilité à 0,5.
— "NOVILON" n'est plus amianté à partir de 1981, donc sa probabilité est 0.
Nous trouvons alors :

$$p_a(RevêtementsMuraux, 1982) = \frac{1 + 0,5 + 0 + 0,5}{4} = 0,5$$

Cependant, l'une des difficultés est que l'INRS et l'ANDEVA se focalisent uniquement sur les produits ayant été amiantés pendant au moins une période durant leur commercialisation. Ne disposant pas du nombre réel de produits commercialisés à une période donnée,

nous estimons que le nombre de produits commercialisés total est largement sous-estimé et ce nombre peut varier en fonction des années. Nous proposons de le réajuster en se basant sur l'ensemble des diagnostics de prélèvement disponibles. Nous comparons pour cela, pour une année donnée d, la proportion de produits amiantés dans les ressources quelle que soit la classe de produit (i.e. $F(p_k) = Produit$) par rapport à l'ensemble des produits commercialisés issus des ressources externes, avec la proportion de produits amiantés dans les diagnostics réalisés pour cette même année par rapport à l'ensemble des diagnostics posés pour l'année. Cela nous permet de déterminer quelle est la proportion α de produits commercialisés manquants que nous considérons comme étant non amiantés.

$$p_a(p_k, d) \times (1 + \omega) = \frac{\sum_{p_j \in F(p_k)} p_{ext}(p_j, d)}{\mid p_j \mid + (\alpha \times \mid p_j \mid)} \tag{2}$$

Dans l'équation 2, ω représente la différence entre la probabilité calculée et la probabilité réelle. Il résulte de la comparaison du ratio de produits amiantés dans les diagnostics avec le ratio de produits amiantés dans les ressources externes (INRS et ANDEVA) de la même année. Il est calculé comme suit :

$$\omega = \frac{\mid p_{diag,a} \mid}{\mid p_{diag} \mid} - \frac{\sum_{p_k \in Produit} p_a(p_k, d)}{\mid p_k \mid}$$

où :
$\mid p_{diag,a} \mid$ est le nombre de produits amiantés dans les diagnostics, et $\mid p_{diag} \mid$ est le nombre total des produits dans les diagnostics.

En utilisant l'équation 2, nous trouvons :

$$\alpha = \frac{\sum_{p_j \in F(p_k)} p_{ext}(p_j, d)}{\mid p_j \mid \times p_a(p_k, d) \times (1 + \omega)} - 1 \Rightarrow \alpha = \frac{-\omega}{1 + \omega} \tag{3}$$

Pour calculer la probabilité de présence d'amiante dans une localisation, une structure ou un bâtiment, nous considérons, comme l'expert, qu'il s'agit de la valeur maximum de l'ensemble des valeurs de probabilité de présence d'amiante des produits p_k qui composent la localisation l_i, puis de l'ensemble des localisations qui participent à une structure et enfin l'ensemble des structures composant le bâtiment. Ainsi, pour une localisation, nous appliquons :

$$p_a(l_i, d) = Max(p_a(p_k, d))$$

De même, pour les structures s_i, il s'agira du maximum des probabilités de ses localisations l_k :

$$p_a(s_i, d) = Max(p_a(l_k, d))$$

Finalement, pour les bâtiments b_i, nous choisissons la valeur maximale des probabilités de ses structures s_k :

$$p_a(b_i, d) = Max(p_a(s_k, d))$$

Pour modéliser les probabilités amiante calculées (figure 4) dans l'ontologie Amiante, nous avons défini le concept de caractéristique calculée qui décrit les caractéristiques Amiante supposées du produit inconnu qui a été utilisé dans un bâtiment. Une instance de ce concept est décrite par la probabilité calculée, l'année de calcul qui va correspondre à l'année du bâtiment et la classe de la probabilité. Seulement deux classes de probabilité ont été définies : forte, et faible (seuil à déterminer expérimentalement). En effet, compte tenu de l'incomplétude des données, une probabilité 1 ou 0 ne nous permet pas de prédire avec certitude que le bâtiment est ou n'est pas amianté.

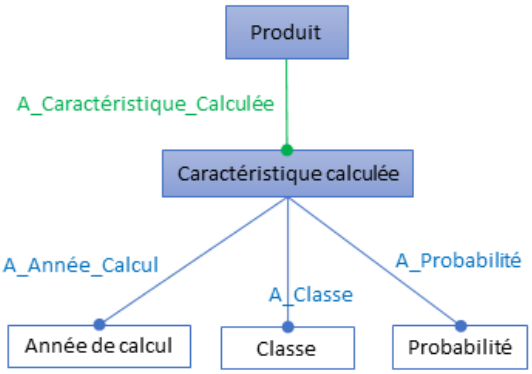


FIGURE *4 – L'ontologie Amiante avec les caractéristiques calculées*

5 Expérimentations

Nous avons mené une première expérimentation qui permet d'enrichir et de peupler l'ontologie Amiante en conservant les données origines et en calculant les caractéristiques des données fusionnées. Nous avons ensuite calculé la probabilité de présence d'amiante dans un bâtiment et de ces éléments pour un ensemble de données synthétiques que nous avons générées. L'objectif est d'observer l'évolution de cette probabilité au fur et à mesure des années sur les résultats obtenus.

5.1 Jeux de données

Description de l'ANDEVA et de l'INRS

Le fichier ANDEVA décrit 650 produits (28,2 Ko) sous forme tabulaire. Chaque produit est décrit par :

— la classe de produit,
— un ensemble de noms de produits qui présentent des propriétés communes et le nom du fournisseur.
— la date à partir de laquelle les produits ne sont plus amiantés.

Le fichier INRS décrit 300 produits (28,1 Ko), sous forme d'un tableau qui contient :

— un ensemble de noms de produits,
— le nom du fournisseur.
— les intervalles de temps où les produits sont amiantés avec un degré d'incertitude.
— le type d'amiante.
— les types d'utilisation (i.e. classes de produit).

Génération de données synthétiques décrivant des bâtiments

Pour éviter d'obtenir des incohérences dans les données synthétiques décrivant les composants du bâtiment (ex. une localisation de type fenêtre ne peut contenir des produits de type ciment), nous avons défini un ensemble de contraintes que le processus de génération doit respecter. Ces 63 contraintes représentent des contraintes de domaine et de co-domaine pour les relations entre les structures et les localisations, et entre les localisations et les classes de produits.

La génération s'effectue de la manière suivante : nous créons un bâtiment dont la date de construction est aléatoirement choisie entre 1946 et 1997. Nous créons pour ce bâtiment un nombre aléatoire de structures que nous associons aléatoirement à une ou plusieurs localisations possibles en respectant les contraintes. De même pour chaque localisation générée nous

associons aléatoirement un ou plusieurs produits. Nous avons ainsi généré cent vingt (120) descriptions de bâtiments qui comportent 1133 produits.

5.2 Analyse quantitative des résultats

Enrichissement et Peuplement de l'ontologie avec l'INRS et l'ANDEVA

L'étape d'extraction de l'information et de fusion des produits commercialisés issus des deux ressources externes nous a permis de créer soixante-quatre (64) sous-classes de produit (e.g. enduit, colle, ...) et six cent quatre-vingt-quatorze (694) instances de produits commercialisés. Presque tous les produits de l'INRS sont également présents dans l'ANDEVA (256/300). Seulement 35 produits conduisent à des valeurs conflictuelles pour les caractéristiques amiantes.

Calcul de la probabilité de présence d'amiante dans les éléments de bâtiment

Nous avons utilisé les diagnostics posés sur 40 produits pour calculer la valeur de α qui permet de réajuster le nombre de produits commercialisés. Les 40 produits diagnostiqués dont nous disposions concernent tous l'année 1963 pour laquelle $\omega \approx -50,19\%$ ce qui nous conduit après la résolution de l'équation 3 en utilisant les données disponibles, à $\alpha \approx 1,0076$.

En utilisant les ressources externes (ANDEVA et INRS), nous avons calculé la probabilité de l'existence de l'amiante pour chaque classe de produits et pour chaque année (de 1946 jusqu'à 1997). Le graphe de la figure 5 montre l'évolution de cette probabilité pour l'ensemble des produits commercialisés. Ce graphe montre que la probabilité de présence d'amiante reste globalement stable jusqu'à 1972, puis elle décroit jusqu'à atteindre 0 en 1997, année où l'amiante est interdit. Sans le réajustement, la proportion de produits amiantés varie de $92,7\%$ à $44,8\%$. Quand le réajustement est appliqué, la proportion maximum de produits amiantés devient $46,17\%$. La figure 6 montre sur les quatre classes de produits présentées en tableau 5 que la probabilité de l'existence de l'amiante diffère d'une classe de produit à l'autre. Par exemple, les adhésifs amiantés sont peu nombreux et se sont désamiantés ou n'ont plus été commercialisés plus rapidement que les trois autres classes de produits présentées en exemple.

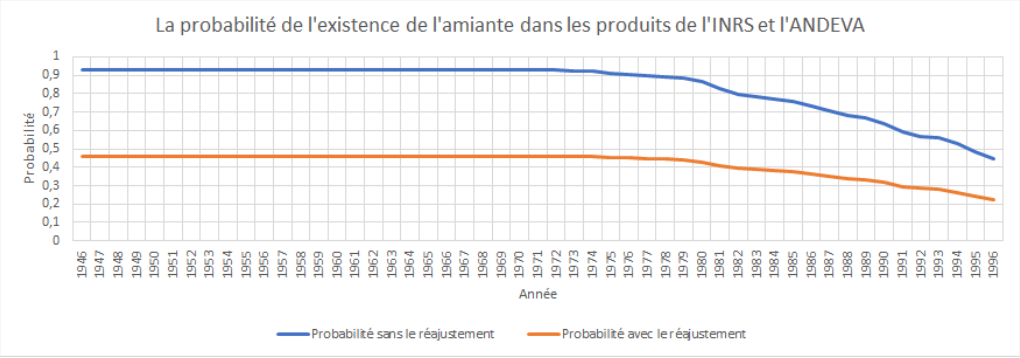


FIGURE *5 – La probabilité de l'existence d'amiante dans les produits de l'INRS et l'ANDEVA en fonction des années*

La figure 7 montre la variation du nombre de bâtiments et leur probabilité moyenne de présence d'amiante en fonction des années dans les 120 descriptions de bâtiments générées. Comme les probabilités de présence d'amiante dans les produits sont propagées sur les éléments de bâtiments puis sur le bâtiment lui-même par une fonction maximum, les résultats ne montrent pas de baisse significative du nombre de bâtiments amiantés sur les données synthétiques. On peut ainsi remarquer que les 4 bâtiments datant de 1947 ont même probabilité moyenne que les 4 bâtiments datant de 1993 (probabilité de 0,2). Il suffit

TABLE 5 – *Caractéristiques des familles de produits choisis*

Famille de produits	Nombre de produits
Adhésif	5
Colles	31
Enduits	19
Mastics	61

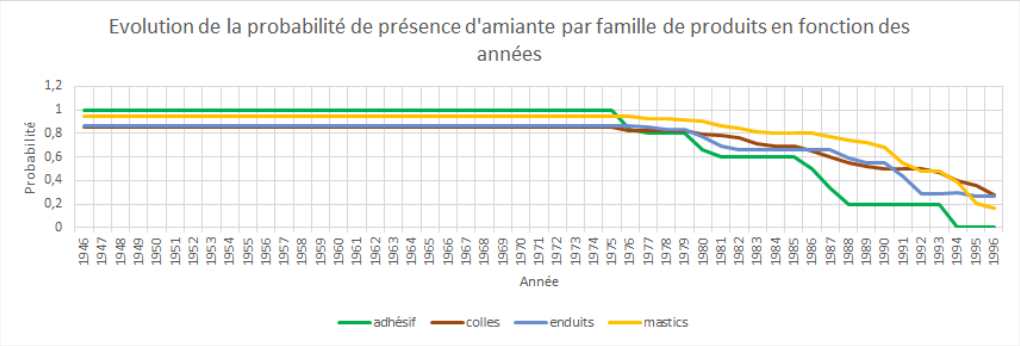


FIGURE 6 – *L'évolution de la probabilité de présence d'amiante par famille de produits en fonction des années*

en effet qu'une classe de produit utilisée soit peu désamiantée pour que le bâtiment entier soit considéré comme ayant un risque assez élevé de présence d'amiante.

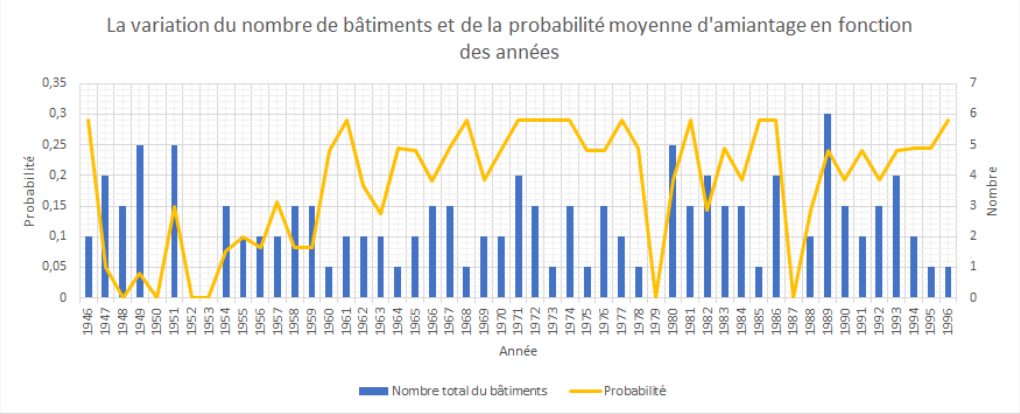


FIGURE 7 – *La variation du nombre de bâtiments et de la probabilité moyenne de présence d'amiante en fonction des années*

5.3 Analyse qualitative des résultats

Nous avons envoyé à l'expert un échantillon de résultats obtenus pour 30 bâtiments choisis aléatoirement pour qu'il valide le processus de calcul et les classes amiante de résultat. A l'origine, nos probabilités se répartissaient en quatre classes amiante (sans amiante, présence possible d'amiante, présence probable d'amiante, présence certaine de l'amiante). L'expert a validé le processus de calcul de probabilité à partir des produits commercialisés. Cependant, son expertise nous a conduit à ne considérer que deux classes amiante (faible et forte) car

l'incomplétude de données ne permet jamais de déduire une présence certaine ou une absence certaine d'amiante mais le seuil à utiliser reste à déterminer en utilisant d'autres diagnostics.

6 Travaux connexes

Approches de prédiction dans les graphes de connaissances

De récentes méthodes d'apprentissage supervisé cherchent à établir des prédictions en utilisant des données graphe (voir Hamilton *et al.* (2017a)) pour un état de l'art. L'objectif peut être de classer le rôle d'une protéine dans un graphe d'interaction biologique (Hamilton *et al.* (2017b)), de prédire le rôle d'une personne dans un réseau social, ou prévoir de nouvelles applications thérapeutiques de molécules de médicament existantes, molécules dont la structure peut être représentée sous forme de graphe. L'une des difficultés est alors de savoir représenter la structure du graphe, ou des éléments statistiques sur cette structure (e.g. degrés, coefficients de clustering) de façon à rendre celui-ci exploitable par des méthodes d'apprentissage. Aussi, certaines approches s'intéressent à l'apprentissage de ces modèles (Hamilton *et al.* (2017a)).
Les approches de programmation logique inductive et de découverte de règles (Lavrac (1994); Galárraga *et al.* (2015); d'Amato *et al.* (2016)) peuvent détecter des règles telles que *"Si une personne est née dans un pays, elle parle probablement la langue de ce pays"*. Ces règles peuvent ensuite être utilisées pour prédire de nouvelles informations et être syntaxiquement guidées par des motifs pour détecter des règles qui concluent sur une information ou une classe d'intérêt pour l'utilisateur (Nebot & Llavori (2012)). De telles règles apprises ont montré de bons résultats pour prédire des valeurs manquantes dans des graphes de connaissances généralistes (Galárraga *et al.* (2015)). Dans notre contexte, nous disposons de trop peu de données d'apprentissage pour apprendre à classifier automatiquement des éléments de bâtiments comme étant amiantés ou non amiantés, que ce soit par des méthodes d'apprentissage profond ou par des méthodes logiques basées sur la détection de règles. L'objectif ici est d'estimer une probabilité de présence d'amiante dans un produit inconnu appartenant à une classe de produit identifiée à partir des propriétés des produits commercialisés utilisés au moment de la construction du bâtiment. Cela revient à découvrir des règles qui, compte tenu de la classe d'un produit et de sa date, impliquent une présence d'amiante dans ce produit avec un certain degré de confiance. Les règles de propagation de cette probabilité à des éléments de bâtiments suivent ensuite une approche pessimiste (utilisation du maximum) qui reproduit le raisonnement d'un expert du bâtiment dans cette analyse de risque.

Incomplétude - Mesure de l'incomplétude

Comme les graphes sont généralement incomplets, les faits manquants ne doivent pas y être considérés comme faux (i.e. Hypothèse du monde ouvert - OWA). La notion de complétude d'un graphe de connaissance ne fait pas toujours sens selon les propriétés et les classes que l'on considère, et sans graphe de référence auquel se comparer (Razniewski *et al.* (2016)). Pour apprendre dans un tel contexte, certaines hypothèses complémentaires ont été introduites (Galárraga *et al.* (2016)). Dans (Galárraga *et al.* (2015)), une hypothèse de complétude partielle a été définie qui propose que, chaque fois qu'au moins un objet pour un sujet et une propriété donnés sont déclarés dans le graphe, tous les objets de cette paire sujet-propriété sont supposés être connus. D'autres approches ont pour objectif de mesurer l'incomplétude des données. Dans (Issa *et al.* (2017)), les auteurs proposent de calculer un schéma idéal à partir des propriétés fréquemment instanciées ensemble et d'en déduire les propriétés qui pourraient être considérées comme manquantes pour les instances de classe. Dans Tanon *et al.* (2018), les auteurs découvrent les cardinalités des propriétés dans les données. Dans notre approche, aucuns des liens entre les produits utilisés et les produits commercialisés ne sont disponibles, et il ne s'agit donc pas d'évaluer l'incomplétude des instances de cette propriété pour un bâtiment donné. Nous proposons d'évaluer le nombre de produits commercialisés non amiantés manquants, et donc le nombre d'instances de classe manquantes, en utilisant le petit nombre de diagnostics disponibles.

Fusion de données

Les approches de fusion de données ont proposé différents critères permettant de gérer les valeurs conflictuelles (Bleiholder & Naumann (2006); Mendes *et al.* (2012)). Certaines stratégies laissent l'utilisateur décider de la meilleure valeur à affecter à l'entité. D'autres stratégies sont automatiques et se basent sur des fonctions de filtrage qui choisissent la valeur la plus récente, la plus fréquente, la plus précise ou la plus fiable quand on dispose de la réputation des sources de données, ou sur des fonctions d'agrégation qui combinent les valeurs possibles en utilisant la moyenne, le minimum ou encore le maximum des valeurs possibles selon l'application. Dans notre approche, les conflits apparaissent lors de la fusion des probabilités de présence d'amiante dans les produits commercialisés associés aux intervalles de temps. Comme l'INRS et l'ANDEVA sont supposés avoir la même fiabilité et que nous ne disposons pas d'un grand nombre de sources qui permettraient d'envisager de se baser sur la fréquence d'une valeur, nous avons choisi de conserver les valeurs d'origine avec leur provenance, mais d'utiliser la fonction maximum dans la description fusionnée afin de tenir compte de ces deux sources pour évaluer de manière pessimiste le risque de présence d'amiante.

Graphes probabilistes

De nombreux travaux de recherche se sont intéressés à la modélisation, au requêtage et à la fouille de graphes probabilistes (Chekol *et al.* (2017); Benferhat *et al.* (2013)). Dans ces modèles, la probabilité peut être associée (1) aux arcs ou aux arêtes et représenter la probabilité d'existence d'un arc ou d'une arête entre deux nœuds du graphe,ou (2)aux nœuds et représenter la probabilité d'existence de ce nœud, ou (3) être associé aux attributs des nœuds ou des arcs (ou arêtes). Dans notre travail, nous ne représentons pas les liens d'identité qui relient potentiellement le produit inconnu utilisé dans un bâtiment à un produit commercialisé de la même classe, car nous considérons que chaque lien d'identité est équiprobable. Nous représentons de manière "réifiée" la probabilité de présence d'amiante dans les produits commercialisés et dans le produit inconnu (en utilisant le concept de caractéristique extraite ou de caractéristique calculée qui permet de lier un produit à une probabilité d'amiantage en conservant les données temporelles et la provenance de l'information pour les produits commercialisés).

7 Conclusion

Dans ce papier, nous avons présenté une première approche de prédiction de la présence d'amiante dans un bâtiment, approche qui devrait permettre d'aider un opérateur de repérage à décider des prélèvements qu'il est nécessaire d'effectuer dans un bâtiment construit à une date donnée. Nous avons tout d'abord défini une ontologie qui modélise les caractéristiques des bâtiments et les diagnostics quand ils existent. Cette ontologie est enrichie par des données temporelles probabilistes sur la présence d'amiante dans les produits commercialisés décrits dans les ressources externes dont on garde la provenance. Nous avons ensuite proposé une méthode pessimiste de calcul des probabilités de présence d'amiante qui se base sur ces données incertaines et incomplètes.

Les premiers résultats montrent que cette probabilité évolue au fur et à mesure des années et qu'elle varie également en fonction des classes de produits utilisées dans un bâtiment.

Dans des travaux futurs, nous allons tester notre solution sur un ensemble de données réelles fourni par le CSTB. Nous planifions également d'utiliser cette probabilité calculée, la description des bâtiments, l'ontologie et un ensemble conséquent de diagnostics pour apprendre plus précisément les caractéristiques des bâtiments, des structures, des localisations et des produits qui peuvent influer sur la présence d'amiante dans les éléments de bâtiments, en utilisant la sémantique de l'ontologie.

IC 2019

Références

BENFERHAT S., KHELLAF F. & ZEDDIGHA I. (2013). A possibilistic graphical model for handling decision problems under uncertainty. In *8th conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-13)* : Atlantis Press.

BLEIHOLDER J. & NAUMANN F. (2006). *Conflict handling strategies in an integrated information system*. Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät ....

CHEKOL M. W., PIRRÒ G., SCHOENFISCH J. & STUCKENSCHMIDT H. (2017). Marrying uncertainty and time in knowledge graphs. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, p. 88–94.

D'AMATO C., TETTAMANZI A. G. B. & TRAN D. M. (2016). Evolutionary discovery of multi-relational association rules from ontological knowledge bases. In *Knowledge Engineering and Knowledge Management - 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings*, p. 113–128.

GALÁRRAGA L., RAZNIEWSKI S., AMARILLI A. & SUCHANEK F. M. (2016). Predicting completeness in knowledge bases. *CoRR*, **abs/1612.05786**.

GALÁRRAGA L., TEFLIOUDI C., HOSE K. & SUCHANEK F. M. (2015). Fast rule mining in ontological knowledge bases with AMIE+. *VLDB Journal*, **24**(6), 707–730.

HAMILTON W. L., YING R. & LESKOVEC J. (2017a). Representation learning on graphs : Methods and applications. *IEEE Data Eng. Bull.*, **40**(3), 52–74.

HAMILTON W. L., YING Z. & LESKOVEC J. (2017b). Inductive representation learning on large graphs. In *Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, p. 1025–1035.

ISSA S., PARIS P. & HAMDI F. (2017). Assessing the completeness evolution of dbpedia : A case study. In *Advances in Conceptual Modeling - ER 2017 Workshops AHA, MoBiD, MREBA, OntoCom, and QMMQ, Valencia, Spain, November 6-9, 2017, Proceedings*, p. 238–247.

LAVRAC N. (1994). Inductive logic programming. In *WLP*, p. 146–160 : Institut für Informatik der Universität Zürich.

MENDES P. N., MÜHLEISEN H. & BIZER C. (2012). Sieve : linked data quality assessment and fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, p. 116–123 : Citeseer.

NEBOT V. & LLAVORI R. B. (2012). Finding association rules in semantic web data. *Knowl.-Based Syst.*, **25**(1), 51–62.

RAZNIEWSKI S., SUCHANEK F. M. & NUTT W. (2016). But what do we actually know ? In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction, AKBC@NAACL-HLT 2016, San Diego, CA, USA, June 17, 2016*, p. 40–44.

TANON T. P., STEPANOVA D., RAZNIEWSKI S., MIRZA P. & WEIKUM G. (2018). Completeness-aware rule learning from knowledge graphs. In *IJCAI*, p. 5339–5343 : ijcai.org.

Patrons linguistiques pour l'extraction de tâches dans des transcriptions de réunions

Namrata Patel[1,2], Mathilde Lannes[3], Camille Pradel[3]

[1] VISEO, 38000 Grenoble, France
namrata.patel@viseo.com

[2] UNIVERSITÉ MONTPELLIER 3 - PAUL VALÉRY, 34090 Montpellier, France
namrata.patel@univ-montp3.fr

[3] SYNAPSE DÉVELOPPEMENT, 31000 Toulouse, France
{mathilde.lannes, camille.pradel}@synapse-fr.com

Résumé : Nous présentons une méthode d'extraction d'informations en deux grandes étapes : (1) analyse morphosyntaxique et annotation sémantique du texte, puis (2) identification de patrons linguistiques par application d'un ensemble de règles sur le texte annoté. Elle est appliquée sur des cas d'usage d'extraction de tâches : des transcriptions de réunions. Une évaluation qualitative manuelle sur un jeu de données réduit montre des résultats encourageants.

Mots-clés : Patrons linguistiques, Extraction d'information, Extraction de tâches.

1 Introduction

Nous présentons dans ce papier une approche de détection automatique de tâches qui s'inscrit dans le cadre de réunions de travail ; un moment bien connu de la plupart d'entre nous dans la vie professionnelle, et qui consomme une bonne partie de notre temps.

Une fois celle-ci terminée, il reste encore un compte rendu à produire, témoignant des principaux aspects abordés lors de la réunion, comme les tâches exprimées, les problèmes rencontrés et les décisions prises. Grâce aux outils dits "Speech-to-text", il est aujourd'hui possible d'enregistrer et ensuite de générer une retranscription textuelle de tout ce qui a été dit lors de la réunion. Un enjeu important est alors d'être capable d'extraire automatiquement de ces retranscriptions textuelles, souvent fortement bruitées, des informations pertinentes, telles que les tâches exprimées par les participants de la réunion.

1.1 Contexte - le projet REUs

Notre approche de détection de tâches s'intègre dans le cadre d'un projet FUI, REUs, composé d'industries et de laboratoires scientifiques. REUs se propose de construire des outils permettant d'améliorer l'efficacité des réunions en préparant des documents en aval : le système REUs est un système complexe qui combine des technologies pour la prise de son, pour la transformation de son en texte, ainsi que des technologies dédiées à la compréhension du texte, dont en particulier la détection de tâches.

Dans ce contexte, nous considérons deux cas d'usage de réunions de type suivi de projets, et nous nous focalisons dans un premier temps sur les réunions en langue française :

— un suivi de réunion de type formation de futurs managers dans le cas d'une école de management,
— un suivi de projets dans le contexte de coaching à des entrepreneurs, avec une société de coworking.

Les difficultés pour construire un tel système, même en se limitant à la détection de tâches, sont multiples. Comment associer le destinataire d'une tâche à l'expression de la tâche dans le contexte de retranscriptions de réunions ? Comment distinguer les actions liées à des tâches de celles évoquées simplement en cours de conversation ? Au delà de ces difficultés sémantiques,

il y a la contrainte des données qui sont bruitées, avec des propos qui ne sont pas toujours structurés, et où plusieurs locuteurs interviennent en même temps.

Nous décrirons dans ce qui suit, l'approche développée par deux partenaires industriels du projet, Synapse Développement et Viseo, dans le but de produire un outil "preuve de concept" pour la détection automatique des tâches.

Plus précisément, il s'agit de combiner la technologie de compréhension automatique de la langue, appelée Machine Reading et développée par Synapse Développement, avec l'expertise de Viseo, spécialisée en extraction d'informations par approches symboliques, afin de réaliser cette première "preuve de concept".

2 Etat de l'art

Dans le contexte du projet REUs, Alizadeh *et al.* (2018) ont mené une première étude expérimentale qui traîte les enjeux scientifiques liés aux extractions d'informations dans les retranscriptions de réunions. Cette étude se focalise sur la génération automatique de compte-rendus de réunions et présente des résultats encourageants pour l'utilisation de techniques existantes, robustes. Elle a été faite sur des corpus existants, en anglais. Notre approche reprend et approfondit cette étude sur un des enjeux scientifiques, l'extraction des tâches évoquées lors de la réunion, et l'évalue sur un corpus français constitué dans le cadre du projet REUs.

Parmi d'autres travaux qui se sont intéressés à l'extraction d'informations dans les retranscriptions de réunions, le système le plus abouti est celui réalisé par Tur *et al.* (2010) appelé CALO. CALO est un système d'analyse de réunions qui retranscrit automatiquement les minutes de la réunion en texte puis identifie et annote ses différentes parties : les thèmes, les tours de parole, les actions, les décisions et fournit deux résumés (abstractif et extractif). En ce qui concerne l'extraction d'actions et de décisions, CALO utilise une approche structurelle. Les différentes prises de paroles de la réunion sont classées en fonction de leur rôle dans le processus : "définition de tâche", "accord" et "acceptation d'une responsabilité". Ensuite, les actions (Purver *et al.*, 2007) et les décisions (Fernández *et al.*, 2008) sont détectées. Cette approche répond très bien aux besoins du projet REUs, mais étant développé pour des réunions en anglais, présente des enjeux linguistiques conséquents dus à la différence structurelle des deux langues. Nous suivons donc une approche structurelle inspirée de celle-ci pour la détection de tâches, en exploitant la structure CDS utilisée dans l'analyse Machine Reading (§3.1).

Dans la littérature en anglais, on retrouve différents termes pour désigner les énoncés des tâches. Paradoxalement, le terme *task* n'est pas aussi employé que, par exemple, *request* ou *action item*. On rencontre également le terme *assignment*. DAMSL (Core, 1998), un formalisme d'annotation de dialogues, définit des catégories d'actes de dialogue pour les actions, les demandes d'information et suggestion. Parmi les travaux récents, Microsoft Research (Chen *et al.*, 2015) ont montré l'utilité de réseaux de neurones à convolution pour la reconnaissance d'un ensemble prédéfini d'actions dans les transcriptions de réunions.

L'ensemble des travaux cités dans cette section ont été menés sur des corpus en anglais. Pour la détection de tâches en français, il n'existe pas, à ce jour, de corpus annoté. Une des contributions du projet REUs sera de produire un tel corpus.

3 Méthodologie

Comme nous l'avons indiqué dans la section précédente, notre approche de détection de tâches suit une démarche purement symbolique, nous permettant ainsi de contourner le manque de ressources françaises de données liées aux retranscriptions de réunions.

Cette approche symbolique nous permet donc de suivre une démarche qui se repose fortement sur (1) la construction grammaticale des phrases et (2) un vocabulaire dédié aux expressions de tâches, constitué manuellement. Notre algorithme de détection de tâches est ainsi constitué de deux étapes distinctes :

1. **Machine Reading :** analyse morphosyntaxique et annotation sémantique du texte.
2. **Détection de tâches :** application d'un ensemble de règles sur le texte annoté permettant d'identifier des patrons linguistiques contenant des expressions de tâches.

La première étape, le Machine Reading, permet d'analyser le texte et de le découper en clauses structurées de type <sujet> <verbe> <complément>, reconstituées à partir des dépendances morphologiques entre le sujet, le verbe et le complément. Ces clauses sont également enrichies d'annotations sémantiques. La deuxième étape prend en entrée les clauses générées par l'étape de Machine Reading, et grâce à la structure et aux annotations de ces clauses, permet de détecter, parmi elles, les clauses qui expriment des tâches. Ces tâches sont identifiées par l'application de règles linguistiques et d'un vocabulaire dédié, élaborés manuellement suite à une analyse préalable de retranscriptions de réunions.

Nous décrivons ces deux étapes plus en détail dans les sous-sections qui suivent.

3.1 Présentation du Machine Reading

Le Machine Reading s'approche d'un composant d'annotation en rôles sémantiques ; l'application du Machine Reading sur un texte produit un ensemble de structures appelées CDS (Clause Description Structure) décrivant chacune une clause. Une clause est une unité lexicale portant sur une formule actancielle ; elle est identifiée dans le texte analysé sous forme de prédicat appliqué à des arguments (sujet, objet, compléments).

L'architecture du système Machine Reading est composée de plusieurs couches effectuant un traitement syntaxique ou sémantique ; elle est détaillée dans les sous-sections qui suivent.

3.1.1 Parsing et désambiguisation lexicale

Nous utilisons un parser interne qui effectue tout d'abord une désambiguïsation lexicale — est-ce un verbe ? un nom ? une préposition ? — et une lemmatisation. Ensuite, le parser sépare les différentes clauses, regroupe les expressions, définit les parties du discours et cherche toutes les fonctions grammaticales (sujet, verbe, objet, direct ou indirect, autres compléments).

Enfin, pour tous les mots polysémiques, un module de désambiguïsation lexicale détecte le sens du mot. Pour la langue française, le taux de réussite est d'environ 87%. Les sens désambiguïsés sont directement liés dans notre taxonomie interne.

3.1.2 Reconnaissance et résolution d'entités nommées

Un détecteur d'entités nommées regroupe les entités nommées. Les entités nommées détectées sont les suivantes : noms de personnes, organisations et localisations, mais également fonctions (directeur, étudiant, etc.), heure (relative ou absolue), nombres, etc. Ces entités sont liées entre elles lorsqu'elles font référence à la même entité, comme par exemple "Toulouse" et "la Ville rose".

3.1.3 Résolution d'anaphores

Nous considérons comme anaphore tous les pronoms personnels (je, moi, lui, son, elle…), tous les pronoms et adjectifs démonstratifs (celui-ci, ceux-ci, ceux-là), tous les pronoms et adjectifs possessifs (les miens, les siens, les nôtres…) et, bien sûr, les pronoms relatifs (qui, lequel, quoi…).

Lors de l'analyse, le système crée un tableau avec tous les référents possibles pour l'anaphore (noms propres, noms communs, expressions, clauses, citations) avec beaucoup d'informations grammaticales et sémantiques comme le genre, le nombre ou encore le type d'entité nommée. Après l'analyse syntaxique et la désambiguïsation des mots, les diverses anaphores sont résolues dans la phrase par comparaison avec le tableau de référents.

Nos résultats à cette étape sont bons, équivalents ou meilleurs que l'état de l'art. Selon le type de pronoms et la langue, le taux de réussite de notre solutionneur d'anaphore est compris entre 62% et 93%.

IC 2018

3.1.4 Relations implicites à explicites

Lorsqu'il y a des sujets ou des objets coordonnés (par exemple, "Papa et Maman"), notre système garde la trace de cette coordination. Par exemple, avec la coordination "Papa et Maman", le système sauvegardera trois CDS différents, un avec le sujet coordonné et deux pour chaque terme du sujet. Le but de cette séparation est de trouver des correspondances possibles avec un seul terme de la coordination. Mais, au-delà de cette décomposition très simple, notre analyseur effectue des opérations plus complexes. Par exemple, dans la phrase "Sarah Watson était un médecin bien connu qui voyageait souvent pour soigner des patients", notre système crée quatre CDS différents en utilisant la réciprocité et en ajoutant des informations implicites. Ce mécanisme existe également pour les structures de CDS, comme décrit dans le paragraphe suivant.

3.1.5 Création et persistance de CDS

Nous décrivons dans cette sous-section les principales caractéristiques des structures de CDS. Nous considérons d'abord l'attribut comme un objet. Les composants principaux de la structure sont les descriptions d'une clause, normalement composée d'un sujet, d'un verbe et d'un objet ou attribut. Bien entendu, la structure autorise de nombreux autres composants, par exemple objet indirect, contexte temporel, contexte spatial... Chaque composant est une sous-structure avec les mots complets, le lemme, les compléments possibles, la préposition éventuelle, les attributs (adjectifs) et ainsi de suite.

Pour les verbes, s'il existe un verbe modal, seul le dernier verbe est pris en compte, mais la relation de modalité est conservée dans la structure. Bien sûr, la négation ou la semi-négation sont également des attributs du verbe dans la structure. Si une forme passive est rencontrée, le sujet réel devient le sujet du CDS et le sujet grammatical devient l'objet. Lorsque le système rencontre un adjectif possessif, un CDS spécifique est créé avec un lien de possession. Par exemple, dans la phrase "Sarah Watson était un médecin bien connu qui voyageait souvent pour soigner des patients", le système créera quatre CDS différents, le premier avec "Sarah Watson" comme sujet, "être" comme verbe et "médecin connu" comme objet. Le second CDS aura "Sarah Watson" en tant que sujet, "voyage" en tant que verbe, "traiter les patients" en tant qu'objet indirect. Le troisième aura "Sarah Watson" en tant que sujet, "traiter" en tant que verbe, "patients" en tant qu'objet et le quatrième CDS aura "docteur" en tant que sujet, "traiter" comme un verbe et "des patients" comme un objet.

De nouveaux CDS sont également créés lorsqu'il existe une relation réciproque. Par exemple, à la question "Pourquoi le père a-t-il conseillé à son fils de se procurer une valise ?", tout le texte concerne la relation entre le père et son fils, de sorte que la conversation est essentielle pour résoudre l'anaphore et répondre aux questions. Le système gère 347 relations de conversation différentes, par exemple les termes classiques "vendre" et "acheter", ou "mari" et "femme", ou "directeur" et "employé", mais aussi des termes géographiques (sud / nord, inférieur / supérieur à ...) et les durées (avant / après, précédent / suivant ...). Pour tous ces liens, deux CDS sont créés.

Les liens entre les CDS sont également enregistrés. D'autres relations comme "cause", "jugement", "opinion" et "autres" sont importantes lorsque le système fait correspondre le CDS du texte au CDS des correspondances possibles. A la fin, après toutes ces extensions, on peut considérer qu'un véritable étiquetage de rôle sémantique est effectué.

Enfin, le système enregistre également les "référents", qui sont des noms propres et communs trouvés dans les phrases, après résolution des anaphores. Ces référents sont particulièrement utiles lorsque le système ne trouve aucune correspondance entre CDS, sachant que les fréquences dans le texte et dans le vocabulaire habituel sont des arguments des structures de référent.

Par exemple, la phrase "On va fixer une réunion demain." donne le CDS illustré en figure 1.

3.2 Détection de tâches à partir de CDS

Nous décrivons maintenant la deuxième étape de notre algorithme d'extraction de tâches : l'application d'un ensemble de règles destinées à identifier les expressions de tâches présentes

```
"offsets": {
    "start": "0",
    "end": "31"
},
"question": false,
"modifiers": {
    "modal": "aller",
    "negation": false,
    "adverbs": []
},
"action": {
    "syntacticType": "VINF",
    "human": false,
    "gender": "undef",
    "number": "undef",
    "groupType": null,
    "namedEntityType": null,
    "core": "fixer",
    "normalized": "fixer"
},
"subject": {
    "syntacticType": "pronoun",
    "human": false,
    "gender": "undef",
    "number": "undef",
    "groupType": "pronominal",
    "namedEntityType": null,
    "core": "On",
    "normalized": "on"
},
"object": {
    "syntacticType": null,
    "human": false,
    "gender": "fem",
    "number": "singular",
    "groupType": "nominal",
    "namedEntityType": null,
    "core": "une réunion",
    "normalized": "un réunion"
},
"time": {
    "syntacticType": "adverb",
    "human": false,
    "gender": "undef",
    "number": "undef",
    "groupType": null,
    "namedEntityType": null,
    "core": "demain",
    "normalized": "demain"
}
```


FIGURE *1 – CDS issu de la phrase "On va fixer une réunion demain."*

dans la transcription d'une réunion.

Cet ensemble de règles se repose sur les structures CDS, générés à l'issue de la première phase présentée en section 3.1, celle du Machine Reading. Les règles ont été conçues afin d'exploiter au maximum les rôles sémantiques identifiés lors de cette dernière.

En définissant une "tâche" elle-même sous une forme structurellement identique au CDS, cet algorithme permet de détecter une tâche, définie sous forme de prédicat. Il se déroule de la façon suivante :

— Identifier des traits linguistiques correspondants aux expressions de tâches dans les tours de parole,
— Repérer les CDS contenant ces termes linguistiques,
— Exploiter la structure et les annotations sémantiques de ces CDS,
— Construire une tâche avec le sujet, le verbe, l'objet et la temporalité liés à la CDS.

Cet algorithme a été élaboré en appliquant la méthologie suivante :

1. Etude linguistique d'expressions de tâches dans un corpus de réunions
2. Représentation formelle d'une "tâche" sous forme de patrons linguistiques
3. Définition d'une correspondance entre les tâches formelles et les structures CDS
4. Développement de règles permettant d'identifier les CDS qui correspondent à des tâches formelles
5. Reconstitution des tâches détectées en langage naturel

Nous décrivons dans ce qui suit, chacune des phases de notre méthodologie.

3.2.1 Etude linguistique

Afin d'aborder l'enjeu de la détection automatique d'expressions de tâches présentes dans les retranscriptions de réunions, nous avons étudié des retranscriptions manuelles d'enregistrements de réunions de type suivi de projet. Ces retranscriptions manuelles font partie d'un corpus qui a été constitué à partir de réunions enregistrées dans le cadre du projet REUs.

Nous avons pu exploiter, lors de cette étude, les résultats obtenus dans le cadre d'une thèse portant sur la détection de tâches dans des emails, réalisée au sein de l'équipe Viseo (Kalitvianski, 2018). Plus précisément, nous avons analysé les différences linguistiques entre les emails et les réunions afin d'adapter (1) le vocabulaire de tâches et (2) les règles de détection de tâches proposés dans cette thèse.

L'objectif principal de notre étude linguistique des retranscriptions manuelles de réunions a donc été de :

1. Identifier des différences linguistiques entre les expressions de tâches dans les emails et celles dans les réunions
2. Adapter le vocabulaire de tâches dans les emails aux tâches dans les réunions.

La figure 2 récapitule nos observations.

En conclusion, (1) le vocabulaire développé dans le contexte des emails reste valable dans le contexte des réunions, il lui manque des termes spécifiques aux réunions tels que "présenter, tester, montrer". (2) La définition d'une tâche dans les emails n'est pas directement applicable au contexte des réunions à cause des différences structurelles entre les expressions textuelles et orales.

Cette étude nous a mené à (1) compléter le vocabulaire avec des termes fréquents trouvés dans le corpus de réunions et (2) définir une représentation formelle d'une tâche exprimée oralement, en s'inspirant de la structure des CDS.

3.2.2 Représentation formelle d'une tâche

Notre étude comparative entre les expressions de tâches dans les emails et les réunions montre que les acteurs impliqués dans les expressions orales ne sont pas explicitement identifiables, contrairement aux expressions dans les emails. Nous nous sommes donc reposés sur

	Emails	**Réunions**
Contenu	Expressions textuelles	Expressions orales
Formulation de phrases	Bien structurée, phrases complètes	Souvent incomplètes, répétitions de mots
Structure	Figée	Conversationnelle, chevauchement des tours de parole
Acteurs	Identification du destinataire et l'assignataire d'une tâche	Participants difficilement distinguables, même dans les retranscriptions manuelles
Vocabulaire de tâches	Expressions d'actions, spécification de divers types de documents (pdf, xls, etc.)	Expressions d'actions, terminologie liée aux réunions (compte-rendu, rendez-vous, suivi, etc.)

FIGURE *2 – Etude linguistique : tâches exprimés dans les emails vs dans les réunions*

la structure des CDS afin de définir une représentation formelle d'expression orale de tâche. La figure suivante présente cette structure.

Patron linguistique	**Sujet du CDS**	**Verbe du CDS**	**Objet du CDS**	**Exemple**
Tâche type 1	pronom personnel ou entité nommée	verbe du vocabulaire de tâches	objet direct de la proposition	Je vais envoyer un mail
Tâche type 2	pronom personnel ou entité nommée	verbe	terme dans le vocabulaire de tâches	On va fixer une réunion demain
Tâche type 3	pronom personnel ou entité nommée	verbe du vocabulaire de tâches	terme dans le vocabulaire de tâches	Je vais partager ce document avec le client

FIGURE *3 – Représentation formelle (CDS) d'une tâche : patrons linguistiques*

3.2.3 Elaboration de règles linguistiques de détection de tâches

Le but de nos règles linguistiques est d'identifier les patrons linguistiques de tâches formelles (voir figure 3) dans le texte analysé. Comme ces patrons sont basés sur des CDS, l'application des règles permet de filtrer l'ensemble des CDS de ce texte. Chaque règle permet donc de spécifier les contraintes définies pour chaque patron linguistique de la figure 3. Par exemple, la règle qui permet de détecter une "Tâche type 1" est (en langage simple) :

Pour chaque CDS :
— *Si le verbe appartient au vocabulaire des tâches*
— *Si le verbe n'est pas au passé*
— *Si le sujet existe, et est "humain" (l'annotation sémantique du CDS permet de distinguer)*
Alors ce CDS est une "tâche type 1"

Une fois que toutes les règles sont appliquées, chaque CDS identifiée comme tâche est reconstituée de la façon suivante :
— « Action » : la tâche avec le sujet entre crochets suivi du verbe lemmatisé et de l'objet
— « Description » : la phrase contenant la tâche
— « EndDate » : s'il existe, l'attribut temporel du CDS détecté par l'analyse MR

Dans le contexte de l'application REUs, nous avons prévu une interface web qui permettrait aux participants de la réunion d'accéder au contenu généré automatiquement par l'application. Il leur est alors possible de corriger les tâches extraites par notre algorithme. Dans

l'élaboration de nos règles, nous avons donc favorisé la possibilité d'avoir des faux positifs (on pourrait les enlever via l'interface web) plutôt que de faux négatifs (qui seraient donc perdus).

4 Evaluation qualitative

Dans le contexte du projet REUs, nous disposons d'un corpus de 30 transcriptions manuelles de captations de réunions. Ces extraits étant tirés de véritables réunions, ils comportent des nuisances sonores ; les bruits ambiants, les hésitations ou encore les coupures de parole des participants rendent difficile l'extraction de tâches.

Afin d'évaluer notre approche, nous utilisons trois transcriptions parlant de sujets différents. La première est une web-conférence sur le suivi d'un projet, la seconde concerne un recrutement au sein d'un espace de co-working et la dernière est un point d'avancement de projet industriel. Nous construisons tout d'abord des documents de référence en extrayant manuellement les tâches de chacune des transcriptions. Nous comparons ensuite ces tâches identifiées manuellement à celles extraites automatiquement par notre système. La comparaison s'effectue à l'aide de trois métriques :

— la précision $p = \frac{\text{nombre de tâches correctes extraites}}{\text{nombre de tâches extraites}}$
— le rappel $r = \frac{\text{nombre de tâches correctes extraites}}{\text{nombre de tâches extraites à la main}}$
— le F-score $f = 2 \cdot \frac{p \cdot r}{p+r}$

Le tableau ci-dessous représente les résultats obtenus pour les trois extraits de réunions.

	Durée	**Nb mots**	**Nb tâches correctes**	**Précision**	**Rappel**	**F-score**
Web-conférence	00 :44 :40	9 768	47	0.296	0.511	0.375
Co-working	00 :24 :12	5 422	7	0.179	0.714	0.286
Projet industriel	00 :20 :46	4 178	35	0,6	0,42	0,5

FIGURE *4 – Evaluation de l'approche*

Ces résultats modestes s'expliquent principalement par le bruit des transcriptions utilisées. En effet, les réunions enregistrées peuvent se dérouler dans des environnements bruyants, à côté d'autres réunions. Nous avons également constaté que les sorties du Machine Reading pâtissent d'une mauvaise interprétation du vocabulaire métier spécifique (notamment des sigles et acronymes). Enfin, il n'est pas rare qu'une même tâche soit répétée plus d'une fois au cours de la réunion ; elle est alors annotée manuellement comme une seule tâche mais extraite à chaque occurence par le composant évalué (qui ne dispose pas de mécanisme d'unification), ce qui a pour conséquence de dégrader la précision.

Il faut également noter que les moins bons résultats ont été obtenus sur une réunion (Co-working, un entretien personnel) s'éloignant clairement du type des réunions ciblées par le projet.

Nous considérons donc ces résultats préliminaires comme encourageants. Une mise en production nécessiterait des efforts supplémentaires, comme la prise en compte du vocabulaire métier et le regroupement de tâches identiques.

Ces efforts permettront d'envisager l'intégration en production du composant d'extraction de tâches dans le système REUs. Cette extraction automatique, disposant d'un bon rappel, devrait être couplée à une validation manuelle pour apporter une fiabilité suffisante : la personne éditant le compte-rendu n'aurait la plupart du temps qu'à reformuler certaines tâches et supprimer les faux positifs, ce qui représente un gain de temps par rapport à une rédaction à partir de rien.

5 Conclusion

Notre approche vise à extraire automatiquement des tâches depuis des enregistrements de réunions afin de préparer des documents de travail en aval. Le système que nous avons mis au point présente des résultats encourageants ; nous identifions plusieurs pistes de travail dans notre analyse linguistique afin d'améliorer ces résultats.

6 Remerciements

Ce travail est soutenu par le projet FUI 22 REUs (N DOS0053568/00).

Références

ALIZADEH P., CELLIER P., CHARNOIS T., CRÉMILLEUX B. & ZIMMERMANN A. (2018). Étude expérimentale d'extraction d'information dans des retranscriptions de réunions. In *Traitement automatique du langage naturel (TALN)*.

CHEN Y.-N., HAKKANI-TÜR D. & HE X. (2015). Detecting actionable items in meetings by convolutional deep structured semantic models. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, p. 375–382 : IEEE.

CORE M. (1998). Analyzing and predicting patterns of damsl utterance tags. In *Proceedings of the AAAI spring symposium on Applying machine learning to discourse processing*.

FERNÁNDEZ R., FRAMPTON M., EHLEN P., PURVER M. & PETERS S. (2008). Modelling and detecting decisions in multi-party dialogue. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, p. 156–163 : Association for Computational Linguistics.

KALITVIANSKI R. (2018). *Traitements formels et sémantiques des échanges et des documents textuels liés à des activités collaboratives*. PhD thesis. Thèse de doctorat dirigée par Boitet, Christian et Bellynck, Valérie Informatique Grenoble Alpes 2018.

PURVER M., DOWDING J., NIEKRASZ J., EHLEN P., NOORBALOOCHI S. & PETERS S. (2007). Detecting and summarizing action items in multi-party dialogue. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, p. 200–211.

TUR G., STOLCKE A., VOSS L., PETERS S., HAKKANI-TUR D., DOWDING J., FAVRE B., FERNÁNDEZ R., FRAMPTON M., FRANDSEN M. *et al.* (2010). The calo meeting assistant system. *IEEE Transactions on Audio, Speech, and Language Processing*, **18**(6), 1601–1611.

Un navigateur pour le Web des données liées

Nicolas Chauvat[1], Fabien Amarger[1], Laurent Wouters[2]

[1] LOGILAB, 104 boulevard Louis-Auguste Blanqui, 75013, Paris
prénom.nom@logilab.fr

[2] CÉNOTÉLIE, 91 rue du Faubourg Saint Honoré, 75008 Paris
lwouters@cenotelie.fr

Résumé : *Le web des documents et le web des données liées utilisent tous les deux les URL pour identifier les ressources et le protocole HTTP pour les échanger. Les navigateurs web actuels sont limités à l'affichage des documents HTML et, nativement, ils ne permettent pas l'affichage de données RDF. Nous proposons d'étendre les fonctionnalités des navigateurs pour parcourir le web dans son intégralité de la manière la plus intuitive possible, c'est-à-dire en suivant les liens entre ressources sans que leur format constitue un obstacle. Pour y parvenir, nous proposons une extension de navigateur permettant de visualiser les ressources RDF et de naviguer sur le Web en passant d'une ressource HTML à une ressource RDF de manière transparente.*

Mots-clés : Web des données liées, Navigateur web, Visualisation, RDF

1 Introduction

L'interconnexion des réseaux informatiques, nommée l'Internet, a permis l'interconnexion des systèmes documentaires et la réalisation de l'idée d'hypertexte sous la forme du Web, un système global d'identification et d'échange d'objets numériques. Le premier âge du Web s'est limité à un Web des documents liés, téléchargés et affichés par un navigateur hypertexte. Avec l'avènement du Web des données, ce sont des données liées qui sont maintenant échangées entre le client et le serveur. Ces données sont transmises de manière structurée, sans être encapsulées dans des documents et des descriptions textuelles. Puisque nous utilisons quotidiennement un navigateur hypertexte pour parcourir le Web des documents, il nous a paru nécessaire de nous interroger sur ce que pourrait être un navigateur pour le Web des données, qui nous permettrait de parcourir le gigantesque graphe global des données liées. Nous avons d'abord constaté que ce changement n'a rien de révolutionnaire : les données comme les documents sont désignées par des URL, téléchargées via HTTP et contiennent des liens vers d'autres données ou d'autres documents. La différence principale tient au fait que les documents sont décrits en HTML alors que les données sont exprimées dans de multiples vocabulaires métiers sur la base du formalisme RDF. En argumentant d'une part que rien ne devrait empêcher de créer des liens hypertextes entre les données et les documents et d'autre part que les documents ne sont qu'un type particulier de données encodées avec un vocabulaire dédié à la mise en page de texte, nous arrivons à la conclusion qu'il n'y a qu'un et un seul Web, et qu'il devrait nous suffire d'un et d'un seul navigateur. Le mécanisme de WebExtension permet justement d'étendre les fonctionnalités des principaux navigateurs (tels que Firefox et Chrome). Nous en avons tiré parti dans le but d'aboutir à une navigation aussi homogène et transparente que possible pour l'utilisateur, qui progresse en suivant les liens présents à l'écran et consulte tantôt des documents HTML, tantôt des données RDF, sans qu'une opération particulière ne soit nécessaire pour afficher les données RDF désignées par l'URL visible dans la barre d'adresse. Ayant constaté qu'un affichage reposant sur la structure du RDF, qu'il s'agisse d'un graphe ou d'une liste de triplets, était peu ergonomique faute de rendre compte de la nature des données, nous avons cherché à adapter la visualisation des données à leur type RDF. Ceci nous a conduit à un avantageux découplage entre le serveur qui fournit les données et le client qui choisit leur mise en page. La navigation en devient plus homogène, puisque les données de même nature peuvent être affichées de la même manière, indépendamment du serveur qui les publie.

Dans la suite de cet article, nous positionnons notre approche par rapport aux travaux qui se sont intéressés à la visualisation des données RDF, puis nous détaillons nos réalisations et nous les illustrons par une démonstration de deux cas concrets de navigation, avant d'ouvrir sur les perspectives que nous envisageons pour la suite de ce projet.

2 État de l'art

Le Web des données ouvertes et liées, aussi nommé Linked Open Data ou LOD, repose sur un ensemble de promesses comprenant :

— rendre les ressources (au sens large) adressables ;
— déréférencer l'adresse d'une ressource donnant accès aux connaissances relatives à cette ressource ;
— pouvoir faire référence à d'autres ressources pour décrire les connaissances associées à une ressource.

Dans une utilisation classique de RDF, ces promesses se déclinent par l'utilisation d'URI pour faire référence à une ressource, le déréférencement d'une telle URI renvoie un document RDF relatif à la ressource et les nœuds URI utilisés dans ces documents peuvent à leur tour être déréférencés, et ainsi de suite. Le LOD repose également sur la mise à disposition des données, en général par le biais d'entrepôts qu'il est possible de requêter. Afin de visualiser et naviguer dans les données du LOD, il est possible de distinguer deux paradigmes différents :

L'approche **centralisée** part du principe que l'intégralité des connaissances disponibles est accessible dans un unique entrepôt de triplets RDF qu'il s'agira de visualiser en tout ou partie. La difficulté est alors souvent de permettre à un utilisateur d'appréhender une nouvelle base de connaissances dont la taille peut être un obstacle à la compréhension.

L'approche **distribuée** part du principe que les connaissances sont réparties sur plusieurs serveurs à travers le web et qu'il va falloir naviguer de l'un à l'autre en suivant les liens pour aboutir au résultat de sa recherche. La difficulté principale est alors qu'on ne sait pas *a priori* où seront hébergées les données, ni quelles seront les ontologies utilisées pour définir les ressources rencontrées. Bien que centrée sur une ressource à la fois, cette approche permet de naviguer à travers l'ensemble des données via les liens présents. C'est en cela que l'analogie entre session de recherche sur le web de documents et session de recherche de connaissances sur le web des données liées, nous paraît pertinente.

Dans cette section nous étudierons les approches existantes pour ces deux paradigmes et nous aborderons une expérience interne à l'entreprise Logilab concernant l'utilisation de vues spécifiques à des vocabulaires RDF. Dans ce but, nous avons réincarné sous une forme plus moderne les idées développées durant les années 2000 dans le cadre de l'interface utilisateur du logiciel CubicWeb [1].

2.1 Consultation d'une base de connaissances

La question de la visualisation et de la navigation pour une base de connaissances complète se pose dans le cas d'entrepôts uniques, par exemple DBPedia. L'enjeu dans de tels cas est alors principalement exploratoire. Les approches relevant de ce paradigme permettent la visualisation et la navigation de ces bases en s'appuyant sur une logique de graphe. On peut ainsi naviguer de proche en proche entre les nœuds et obtenir une visualisation de ceux-ci, en général spécifique au fournisseur de l'entrepôt. Les travaux tels que IsaVizPietriga (2003), V-OWLLohmann *et al.* (2016), ou encore VizSKOSDestandau (2016) pour la visualisation de vocabulaire SKOS, proposent des visualisations génériques pour une exploration du graphe de connaissances dans son intégralité. Ces approches sont particulièrement pertinentes pour une visualisation d'une base de connaissances dans son intégralité et leur exploration sans connaître au préalable, ni le vocabulaire, ni ce que l'utilisateur cherche. Cependant, leur utilisation n'est possible que lorsque l'intégralité de la base de connaissance est accessible localement. Comme présenté dans l'introduction 1 nous souhaitons proposer un navigateur pour

1. `https://cubicweb.org`

le web des données liées. Il n'est donc pas envisageable de stocker l'intégralité du LOD et c'est pourquoi notre hypothèse initiale est l'affichage spécifique d'une ressource.

2.2 Affichage d'une ressource

Le second paradigme concerne la visualisation d'une ressource particulière. Dans le cadre du LOD, ce paradigme s'intéresse à la visualisation des connaissances à propos d'une ressource telle qu'accessible en déréférençant l'adresse de celle-ci. Dans le cas précis de RDF, cela correspond à s'intéresser à la visualisation spécifique de jeux de données RDF *relativement petits* puisque relatifs à une seule ressource à la fois, contrairement au paradigme précédent. Les approches comprises dans ce paradigme peuvent être catégorisées en deux groupes : les vues génériques et les vues spécifiques.

2.2.1 Vues génériques

Les connaissances sur le web des données liées étant représentées par des triplets RDF, il est possible de proposer une visualisation générique qui permet l'affichage d'une ressource sous la forme de triplets "sujet", "prédicat", "objet". Nous pouvons par exemple citer TabulatorBerners-Lee *et al.* (2008) qui propose un affichage sous forme d'arbre afin de parcourir le chemin de relation pouvant mener vers de nouvelles connaissances. Par exemple il est possible d'afficher les triplets dont une ressource est le sujet, et ensuite de parcourir (récursivement) les triplets dont un objet devient le sujet. La limite des approches proposant des vues génériques est aussi leur force. Bien que les vues n'ont besoin d'aucune configuration ou de développement pour être utilisables directement par un utilisateur, ces approches deviennent vite limitées dans le cadre d'un projet industriel. Effectivement, lorsqu'un client d'une entreprise décide d'avoir un affichage pour ses données exprimées en RDF, il est rare qu'une représentation aussi générique qu'un arbre ou un tableau de triplets soit suffisant, car l'étalon reste les interfaces utilisateur riches des applications faites sur-mesure. Nous pouvons prendre pour exemple une expérience observée chez Logilab. Notre entreprise développe le cadriciel de développement web CubicWeb [2]Simon *et al.* (2013), dont le modèle de données s'approche d'une base de connaissances de part l'utilisation d'entités reliées entre elles par des relations nommées et contenant un certain nombre d'attributs. Afin de visualiser ces données, CubicWeb possèdent un système de génération automatique de vues pour chaque type d'entité et de relation. De cette façon si un nouvel attribut, ou relation, est ajouté à l'entité, alors cette modification est directement visible sur l'interface, que ce soit en visualisation ou en modification. Le client observe donc un changement immédiat et peut émettre des retours. Cependant cette vue, auto générée, est rarement suffisante pour le client. Avec l'explosion des applications complexes côté clients et des technologies telles que React [3], Angular [4] ou encore Vue.js [5] (pour ne citer qu'eux), il devient difficile de proposer la même qualité de visualisation de données avec des vues auto générées et leur configuration devient problématique. Nous observons, dans la majorité de nos projets, la suppression de ces vues auto générées au profit de vues qui sont développées spécifiquement afin de répondre aux besoins des clients. Les vues auto générées ne sont donc pas une solution adaptée à nos besoins.

2.2.2 Vues spécifiques

Il faut pouvoir proposer des vues spécifiques pour des ressources RDF déréférencées. Pour cela, certains travaux proposent des approches intéressantes. Nous pouvons citer HaystackQuan *et al.* (2003) qui propose la définition de vues spécifiques pour chaque ressource. Bien que cette approche soit particulièrement intéressante, ces vues sont générées côté serveur. Il devient donc nécessaire d'avoir un serveur spécifique qui va servir les ressources et

2. https://www.cubicweb.org/
3. http://reactjs.org/
4. https://angular.io/
5. http://vuejs.org/

les vues associées pour obtenir le résultat souhaité. Notre objectif ici est de proposer une navigation sur le web de données liées, tout autant que sur le web des documents, Quel que soit le serveur fournissant la ressource et le type de ressource. Si chaque serveur proposant une ressource disponible sur le web des données liées devait être dans l'obligation de fournir une vue dédiée à cette ressource, alors la publication de données serait difficilement accessible. De plus cela contredirait la philosophie du web des données liées qui préconise de rendre disponible des données ouvertes et liées quelle que soit leur utilisation. D'autres approches, comme STTLCorby & Zucker (2015) proposent une visualisation spécifique, non pas d'une ressource, mais d'un résultat d'une requête SPARQL. Ce type d'approche ne permet pas une visualisation d'une ressource déréférencée : la ressource désirée étant un résultat d'une requête SPARQL, il faut obligatoirement un serveur SPARQL pour que STTL fonctionne. De plus cela empêche la navigation d'une ressource à une autre alors que le web de documents le permet. Enfin, Piggy BankHuynh *et al.* (2007) propose une extension aux navigateurs traditionnels (Firefox et Chrome) pour permettre l'extension de la visualisation de document grâce à du contenu RDF ou RDFa. Cette approche se concentre sur l'amélioration de l'affichage de document et non pas sur la visualisation de ressources RDF à proprement parler. Pour cela les documents HTML sont parcourus afin d'en extraire un certain nombre de triplets RDF (grâce au micro format ou au "scrapping") pour étendre la visualisation. Néanmoins cette approche est intéressante grâce à la décision d'utiliser une extension de navigateur pour étendre ses fonctionnalités.

2.3 Sélection d'une vue

Nous souhaitons une approche permettant une visualisation d'une ressource RDF déréférencée grâce à une vue spécifique pour un type de ressource donné. Nous réfléchissons donc au moyen de sélectionner une vue qui corresponde à ce que désire l'utilisateur. Les travaux effectués pour FresnelPietriga *et al.* (2006) proposent une réponse à cette problématique et ont bénéficié d'une standardisation du W3C [6] qui se décompose en deux parties. La première partie est FSL [7] un langage de sélection de triplets RDF, qui s'apparente à XPath. De cette manière, il est possible d'exprimer une contrainte de sélection pour l'application d'une vue spécifique, *e.g.* un sélecteur pour les ressources de type $foaf : Person$ ayant une valeur spécifique pour un attribut donné. L'objectif de FSL est de permettre la définition d'un sélecteur de vue et son partage entre les systèmes de visualisation de vue RDF. De cette manière, il devient plus aisé de définir un sélecteur de vue de façon générique et de le réutiliser. La deuxième partie de Fresnel est le formatage d'une ressource. Il ne s'agit pas de vue à proprement parler mais plus de contraintes de formatage telles qu'elles peuvent être définies dans un fichier CSS. Par exemple, nous pouvons spécifier que, pour un sélecteur FSL d'une ressource de type $foaf : Person$, le nom $foaf : name$ de cette ressource sera affiché en gras. Ces contraintes sont, elles aussi, partageables et génériques puisque c'est l'objectif premier de Fresnel. Bien que notre approche n'utilise pas encore le FSL comme un moyen d'exprimer la sélection d'une vue, ce langage nous semble être une approche pertinente. Les contraintes d'affichage peuvent elles aussi être utiles pour s'assurer d'un affichage particulier quelle que soit la vue sélectionnée dans notre approche. Nous reviendrons sur ce point dans cet article. La problématique de sélection de la vue à utiliser pour l'affichage d'une ressource spécifique n'est pas anodine. Nous pensons que, à l'instar du principe selon lequel les données personnelles appartiennent à l'utilisateur proposé dans SolidMansour *et al.* (2016), la décision de la vue à utiliser est de la responsabilité de l'utilisateur lui-même. Pour cela nous proposons un système de serveur de vue configurable que nous détaillerons dans cet article.

Nous avons vu dans ce chapitre que les approches proposant une visualisation d'une base de connaissances entière ne correspondent pas à notre besoin. De plus, les approches de visualisation d'une ressource spécifique ne permettent pas la définition d'une vue spécifique,

6. `https://www.w3.org/2005/04/fresnel-info/`
7. Fresnel Selector Language

générée côté client et quel que soit le serveur distribuant cette ressource. C'est pour cela que nous proposons ici un navigateur pour le web des données liées répondant à ces besoins.

3 Proposition scientifique

Notre approche, présentée dans cette section, a pour objectif de proposer un navigateur pour le web des données liées. Elle s'inscrit donc dans le second paradigme de visualisation présenté dans la section 2. Nous nous intéressons à la possibilité de naviguer entre les ressources du LOD tout en visualisant celles-ci à l'aide de vues spécialisées par catégories de données, mais homogènes quant à leur source, laissant ainsi un contrôle fin de leur visualisation à l'utilisateur final.

3.1 Une extension des navigateurs traditionnels

Notre objectif est de permettre une exploration du web des données liées et du web des documents de manière totalement transparente pour l'utilisateur. C'est-à-dire que nous souhaitons pouvoir passer d'une ressource HTML à une ressource RDF aussi simplement que par un clic sur un lien. Cette transparence va permettre aux utilisateurs d'utiliser les ressources disponibles sur le web des données lées aussi facilement que les ressources du web des documents. Il est alors nécessaire de générer une visualisation HTML à partir de ressources RDF. Compte tenu des règles de sécurité dans les navigateurs, il n'est pas possible de servir une application à une URL qui consulte des données issues d'un autre serveur sans la coopération de ce dernier (règles CORS). Contourner le problème via un serveur proxy poserait des problèmes de passage à l'échelle et de vie privée (espionner trafic). C'est pourquoi, le plus simple est d'étendre les fonctionnalités du navigateur via une Web Extension. Nous avons donc développé et rendu disponible une extension à ces adresses :

Firefox :
— `https://addons.mozilla.org/fr/firefox/addon/linked-data-browser/`
Chrome :
— `https://chrome.google.com/webstore/detail/cubicweb-linked-data-brow/nbhcjnnbhnmbahlofnbfekjaohgmlfcb`

Nous vous encourageons à télécharger et à tester ces extensions et surtout à ne pas hésiter à nous faire des retours pour des améliorations possibles.

3.1.1 Négociation de contenu

Une fois l'extension installée sur le navigateur, chaque page consultée est analysée pour observer si cette ressource (ici documentaire) est disponible sous forme d'un jeu de données RDF. À chaque chargement d'une page, deux méthodes sont utilisées pour détecter si la ressource est aussi disponible en RDF. La première méthode est l'analyse de l'entête de la réponse HTTP afin d'observer si un attribut "LINK" est présent avec une valeur pointant sur une ressource RDF. Cet attribut est défini dans la RFC5988 [8]. La deuxième méthode est de regarder le contenu de la balise "<head>" du contenu HTML, si la balise "<link rel='alternate' >" est disponible avec un type assimilé à une ressource RDF [9]. Nous ne prétendons pas couvrir tous les cas possibles de spécification de l'existence d'une ressource sous un format alternatif, mais ces deux méthodes nous semblent fiables et couramment utilisées. Nous l'avons, par exemple, utilisé sur data.bnf.fr ou encore sur dbpedia.org.

3.1.2 Activation de l'extension

Si au moins une de ces deux conditions est respectée alors l'extension s'active, et une icône apparaît à coté de la barre d'adresse. Si l'utilisateur clique sur cette icône, alors une nouvelle

8. `https://tools.ietf.org/html/rfc5988#section-3`
9. application/rdf+xml, text/turtle, ...

requête HTTP est envoyée en modifiant, dans l'entête HTTP, la valeur du champ "Accept" [10] pour demander le jeu de données RDF associé. De cette manière, l'extension récupère les triplets RDF associés à la ressource et peut ensuite utiliser le système de sélection de vue pour générer une visualisation adaptée. L'extension utilise un ensemble d'heuristiques pour déterminer quelle est l'entité RDF *principale* dans le jeu de données RDF récupéré. Il s'agit dans les cas simples de l'URI à l'origine de l'activation, laquelle peut être différente de l'URI déréférencée pour récupérer le jeu de données. Mais il est tout à fait possible que le sujet principal ne soit pas désigné par l'URI dans la barre d'adresse. C'est notamment le cas lors de l'utilisation de la relation $foaf : focus$ [11].

3.2 Sélection de vue

Une fois une ressource RDF identifiée, le jeu de données associé récupéré et l'URI de l'entité principale définie, l'extension doit sélectionner une vue pour générer la visualisation correspondante. Afin de permettre la diffusion et le contrôle des vues à utiliser, nous proposons la notion de serveur de vues. Ces serveurs rendent accessibles un ensemble de vues et leurs descriptions afin que l'extension puisse identifier quelle vue utiliser pour une ressource donnée.

3.2.1 Serveur de vues

Un serveur de vues met à disposition un ensemble de vues. Pour cela, le point d'entrée est un fichier JSON contenant le descriptif de toutes les vues disponibles sur ce serveur. L'URL de ce fichier JSON est la référence du serveur de vues. L'extension dispose d'une interface de configuration dans laquelle il est possible d'ajouter ou de supprimer un serveur de vue. De cette manière, l'utilisateur contrôle directement les vues qu'il souhaite utiliser ou non. Ce n'est plus le fournisseur des données qui a la responsabilité de la visualisation mais un développeur de vue qui les rendra disponible directement sur ce serveur de vues. De cette manière, la séparation entre la logique métier et l'affichage lors d'un développement d'une application web est d'autant plus respectée, grâce à cette gestion de serveurs de vues. La figure 1 présente un exemple d'un fichier JSON définissant un ensemble de vues.

Ce fichier contient deux vues. Une pour afficher une personne ($foaf : Person$) et une autre pour afficher un livre ($purl : Book$). Chaque vue a les caractéristiques suivantes :

identifier : l'identifiant de la vue

name : le nom de la vue qui sera affichée pour l'utilisateur

description : une description pour avoir un peu plus de détail sur la vue

entrypoint : le nom de l'objet Javascript servant de point d'entrée pour la vue dans la ressource principale

resourceCss : la liste des ressources CSS à importer

resourceJs : la liste des ressources Javascript à importer

resourceMain : la ressource Javascript principale qui contient le point d'entrée

Nous remarquons dans ce fichier de configuration que les deux vues sont disponibles sur le même ordinateur que le client ("http ://localhost :8080/..."). Il est tout à fait possible de distribuer des vues qui ne sont pas sur le même serveur.

3.2.2 Description des différentes vues disponibles

Grâce à ces serveurs de vues, il est possible de distribuer un certain nombre de vues depuis un serveur donné. Il est aussi possible que les vues dont l'utilisateur a besoin se trouvent sur plusieurs serveurs de vues. Par exemple, l'utilisateur souhaite une vue pour $foaf : Person$

10. `https://developer.mozilla.org/fr/docs/Web/HTTP/Headers/Accept`
11. `http://xmlns.com/foaf/spec/#term_focus`

```
[
  {
    "identifier": "::Logilab::Person",
    "name": "Logilab: Person View",
    "description": "Renders a person from a (possibly FoaF) dataset",
    "entrypoint": "VIEW_PERSON_ENTRYPOINT",
    "resourceCss": [
      {
        "location": "remote",
        "uri": "https://stackpath.bootstrapcdn.com/bootstrap/4.1.3/css/bootstrap.min.css"
      }
    ],
    "resourceJs": [],
    "resourceMain": {
      "location": "remote",
      "uri": "http://localhost:8080/view_person.js"
    }
  },
  {
    "identifier": "::Logilab::Book",
    "name": "Logilab: Book View",
    "description": "Renders a book from a dataset",
    "entrypoint": "VIEW_BOOK_ENTRYPOINT",
    "resourceCss": [
      {
        "location": "remote",
        "uri": "https://stackpath.bootstrapcdn.com/bootstrap/4.1.3/css/bootstrap.min.css"
      }
    ],
    "resourceJs": [],
    "resourceMain": {
      "location": "remote",
      "uri": "http://localhost:8080/view_book.js"
    }
  },
]
```


FIGURE *1 – Fichier de configuration d'un serveur de vue*

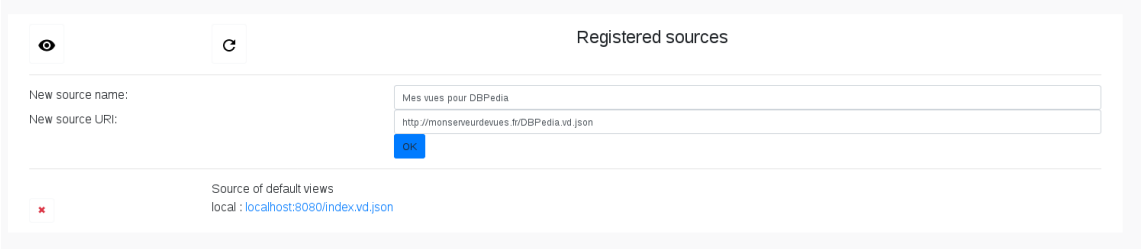


FIGURE *2 – Configuration d'un nouveau serveur de vues dans l'extension*

et une autre vue pour *dbpedia* : *Event* mais ces deux vues ne sont pas définies sur le même serveur. Dans ce cas-là, il est possible de définir plusieurs serveurs de vues comme configuration de l'extension comme illustré sur la figure 2. Sur cette figure un nouveau serveur de vues nommé "Mes vues pour DBPedia" et hébergé sur "http ://monserveurde-vues.fr/DBPedia.vd.json" est en train d'être ajouté. Lorsque le bouton "OK" sera cliqué, les vues de ce nouveau serveur apparaîtront dans la liste des vues disponibles. À chaque ajout d'un nouveau serveur, la liste des vues disponibles sur ce serveur est enregistrée. Cette liste est ensuite utilisée comme vue potentielle pour une ressource donnée.

3.2.3 Algorithme de sélection d'une vue en fonction de la source d'intérêt

Nous venons de voir qu'il est possible de définir plusieurs vues sur un serveur et plusieurs serveurs dans l'extension. Chaque vue dispose d'une fonction "priorityFor" [12] qui retourne un entier pour définir la priorité de cette vue pour visualiser une ressource en particulier. De cette manière dès qu'une ressource est à visualiser par l'extension, celle-ci parcourt toutes les vues disponibles dans les différents serveurs de vues et calcule cette priorité. La vue qui retourne la priorité la plus haute sera alors sélectionnée pour visualiser la ressource. Cette fonction de priorité peut implémenter différents types d'algorithme pour déterminer la pertinence d'une vue pour une ressource en particulier. De cette manière toutes les contraintes peuvent être exprimées ici. Le plus courant reste tout de même un filtre sur le type de la ressource. Par

12. La structure d'une vue sera décrite dans la section 3.3

exemple si l'utilisateur souhaite visualiser une ressource de type $foaf : Person$ la vue identifiée par " : :Logilab : :Person" va probablement retourner une valeur haute pour la priorité. Cette vue regarde le type de la ressource (un triplet de type : "<ressource> rdf :type foaf :Person"), si ce type est bien celui d'une personne du vocabulaire $foaf$, alors elle retourne une valeur forte (dans notre exemple la valeur est 10). Ce procédé a l'avantage d'être relativement simple à implémenter et permet d'avoir un système de sélection de vue à moindre coût. Néanmoins la définition d'une valeur comme priorité est forcément corrélée aux valeurs de priorité qu'ont pu définir les autres vues. Donc une valeur de priorité n'a de sens que dans le contexte d'un ensemble de vues. À partir du moment où les vues sont développées par plusieurs personnes, ces valeurs de priorités deviennent moins pertinentes puisque chaque personne peut définir une valeur arbitrairement. Une évolution intéressante ici serait de prendre en considération les FSL [13] pour définir les contraintes de sélection de vue de manière standardisée et permettre un calcul de priorité plus fin. Une fois une vue sélectionnée, l'utilisateur a toujours la possibilité de changer la vue choisie grâce à l'interface de l'extension. De cette manière si la sélection de la vue n'a pas sélectionné la vue la plus adaptée, l'utilisateur peut décider d'en utiliser une autre.

3.3 Définition d'une vue

Une fois une vue pertinente sélectionnée pour une ressource donnée, la vue est utilisée pour générer la visualisation. Pour cela, chaque vue définit une fonction "*render*" qui sera utilisée pour générer le contenu HTML de la visualisation en fonction de la ressource pertinente. Comme vu dans la section 3.2.1 des ressources externes sont chargées (comme des fichiers CSS ou Javascript) qui seront utilisées ensuite dans la visualisation elle-même. Une ressource peut être liée, par l'intermédiaire de relation, à d'autres ressources déréférencées. L'extension permet de passer d'une ressource à une autre par un simple clic.

3.3.1 Sélection des paramètres d'intérêts pour la vue

Seulement une partie des attributs présents dans le RDF sont pertinents pour une vue donnée. C'est pourquoi dans cette fonction *render*, les attributs d'intérêts sont pré-chargés. Cela signifie qu'un ensemble d'attributs et de relations pertinents pour cette visualisation sont pré-chargés en explorant les triplets du jeu de données retourné par la requête HTTP présentée en 3.1.2. Pour les labels qui sont pertinents, une langue par défaut est configurée dans la vue pour obtenir les chaînes de caractères dans la bonne langue. L'ensemble de ces valeurs sont ensuite fournies à la fonction *render*, qui est chargée de produire le HTML correspondant. L'avantage majeur de cette implémentation réside dans la liberté d'utilisation de ces valeurs et dans le développement de la visualisation. En effet si une personne désire développer une vue en utilisant une technologie spécifique, par exemple React, alors le template de la vue peut être un composant React. Il n'y a, ici, aucune contrainte concernant une technologie plutôt qu'une autre, si ce n'est le javascript initial pour charger la ressource et initialiser le template. Ce n'est pas anodin puisque dans l'univers du développement web, et particulièrement le web appelé "frontend", les technologies évoluent à une vitesse impressionnante. Il devient donc primordial de pouvoir s'adapter à cette évolution. Dans ce cadre, il est tout à fait possible d'utiliser, par exemple, Fresnel afin de réaliser le rendu d'une vue.

3.3.2 Pré-chargement des ressources liées

Dans cette approche, les attributs et relations d'une entité RDF sont très souvent utilisés dans la visualisation de l'entité RDF. Les attributs permettent d'observer directement l'information importante à afficher puisque l'objet du triplet est directement la valeur sous la forme d'une chaîne de caractères ou autre. Néanmoins, pour les relations, la problématique est que l'objet des triplets est une URI référençant une nouvelle entité. L'extension permet d'utiliser les triplets relatifs à cette nouvelle entité, s'ils sont déjà présents dans le jeu de

13. Fresnel Selector Language `https://www.w3.org/2005/04/fresnel-info/fsl/`

données RDF courant, ou bien s'ils peuvent être déréférencés via leur URI en se basant sur les promesses du LOD et pour récupérer des triplets distants. Ces nouveaux triplets peuvent alors être utilisés dans la visualisation de l'entité RDF originale. Un exemple simple est l'affichage d'une vue pour les ressources de type "$foaf : Person$", qui détiennent une relation "$foaf : knows$" vers une autre "$foaf : Person$". Cette autre "$foaf : Person$" est une URI alors que pour la visualisation nous aimerions avoir le nom de cette personne pour afficher la liste des connaissances d'une personne directement sous la forme d'une liste de noms au lieu d'une liste d'URIs. Le choix qui a été fait ici est d'afficher dans un premier temps les URI de ces personnes dans la visualisation, et de récupérer en tâche de fond les noms de ces personnes de manière asynchrone comme décrit ci-dessus. Ainsi, dès que les triplets concernant une personne connue sont récupérés, le nom est extrait et vient remplacer l'URI qui est affichée de manière temporaire. De cette manière, l'interface n'est jamais bloquée ou incomplète. L'information qui y est affichée se précise au fur et à mesure que les connaissances sont récupérées.

3.3.3 Changement de vue dynamique

L'affichage d'une relation revient donc à afficher une URI, puis un label préféré (par exemple le nom d'une personne) dès qu'il est récupéré. Seulement l'URI qui fait référence à cette ressource liée peut, elle aussi, être utilisée comme ressource principale. Si l'utilisateur clique sur une URI déréférencée avec négociation de contenu disponible (comme vue dans la section 3.1.1), alors tout le processus est relancé depuis le début. C'est-à-dire que l'url change dans la barre d'adresse du navigateur, que cette ressource devient la ressource principale, les vues disponibles dans l'ensemble des serveurs de vues sont parcourues pour déterminer la vue la plus pertinente, les attributs et relations pertinents sont récupérés et la visualisation est générée. De cette manière il devient aisé de naviguer d'une ressource à une autre. Un avantage majeur ici est que ce système s'applique quel que soit le serveur qui fournit la ressource. De cette manière, la navigation peut commencer sur notre propre base de connaissances en observant une visualisation d'une personne qui connait "Claude Debussy" qui est référencé par son URI sur DBPedia [14]. Un simple clic sur un lien vers cette URI permet d'afficher une visualisation pour la ressource "$dbpedia : Claude\_Debussy$". Pour aller encore plus loin dans l'interopérabilité et dans la généricité de la navigation, si un lien présent sur une visualisation pointe vers un document, et non plus sur une ressource déréférencée, alors l'extension se désactive automatiquement et laisse place à la visualisation "classique" d'un document HTML par le navigateur. L'hypothèse initiale qui stipulait que nous souhaitons un navigateur qui soit transparent pour l'utilisateur quant au type de la ressource à afficher est ici respectée. Une critique majeure peut être faite sur la nécessité d'installer une extension pour permettre l'affichage de ressources RDF. Il serait donc intéressant d'introduire directement ce fonctionnement dans les navigateurs traditionnels. Si ce fonctionnement convainc une communauté suffisamment grande il pourrait être intéressant de proposer une standardisation de cette forme de communication. Pour l'instant nous nous positionnons comme une preuve de concept. Néanmoins, nous envisageons d'introduire ce type de fonctionnement dans le framework web développé par Logilab [15] pour faciliter le développement de vues dans nos différents projets.

L'ensemble des propositions présentées dans cette section est disponible sous licence LGPL3 à cette adresse : `https://www.cubicweb.org/project/cubicweb-linked-data-browser/`. Une documentation est aussi mise à disposition pour donner plus de détails sur l'implémentation et aider les personnes désirant mettre en œuvre cette approche. Nous restons bien entendu à l'écoute des retours qui peuvent être faits pour améliorer ce système.

14. `http://dbpedia.org/page/Claude_Debussy`
15. CubicWeb`http://www.cubicweb.org`

FIGURE *3 – Capture d'écran de la page HTML concernant Claude Debussy sur DBPedia*

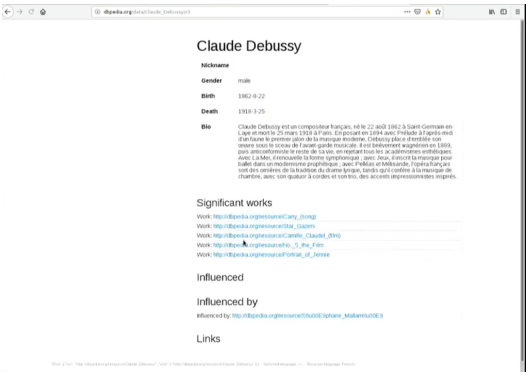


FIGURE *4 – Visualisation générée par la vue " : :Logilab : :Person" avant le chargement des ressources liées*

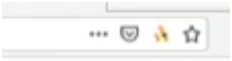


FIGURE *5 – Icône d'activation de l'extension*

4 Démonstration

Afin d'illustrer notre proposition nous allons, dans cette section, présenter deux cas concrets d'utilisation de notre extension. Le premier est une recherche sur le compositeur Claude Debussy pendant laquelle nous passerons d'une visualisation d'un document HTML à la visualisation de données RDF. Le second est l'utilisation de notre extension afin de visualiser les données du site de la conférence SemWeb.pro.

4.1 Claude Debussy sur DBPedia

Notre premier exemple se place suite à une recherche sur votre moteur de recherche préféré des mots clefs "Claude Debussy". Cette session de recherche vous amène directement sur la page HTML concernant la personne "Claude Debussy" sur DBPedia : `http://dbpedia.org/page/Claude_Debussy`. La figure 3 présente cette page.

Sur cette image nous pouvons observer que la visualisation proposée par DBPedia est une liste de triplets dont "$dbpedia : Claude\_Debussy$" est le sujet. Cette visualisation n'est pas modifiable puisque générée par le serveur sous forme HTML. Néanmoins, DBPedia propose la possibilité d'effectuer de la négociation de contenu pour obtenir les données RDF qui ont été utilisées pour générer cette visualisation. Il est possible de le vérifier en regardant dans la barre d'adresse, l'icône de l'extension apparaît comme sur la figure 5. Cette icône n'apparaît que lorsque la négociation de contenu est disponible.

Si nous cliquons sur cette icône, alors l'extension s'active. La première tâche de l'extension est de récupérer les données RDF associées à la page qui était en cours de consultation. Une fois la liste des triplets RDF récupérée, celle-ci est analysée pour déterminer la vue la plus appropriée. Dans notre cas, le triplet indiquant que la ressource "$dbpedia : Claude\_Debussy$" est de type "$foaf : Person$" permet de sélectionner la vue " : :Logilab : :Person" fournie par le serveur dont la configuration est présentée sur la figure 1 à la page 7. Lorsque l'extension a sélectionné la vue la plus appropriée (ici " : :Logilab : :Person"), celle-ci est utilisée afin de générer la visualisation pour les données RDF récupérées. La figure 4 permet de voir la visualisation générée. Cette visualisation est relativement simple, néanmoins, il est possible de voir que seulement certaines données concernant Claude De-

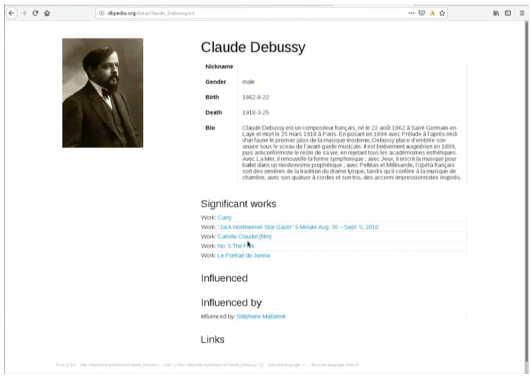


FIGURE 6 – *Visualisation générée par la vue " : :Logilab : :Person" avec les données des ressources liées*

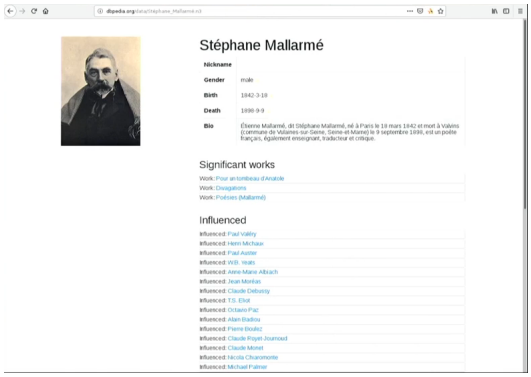


FIGURE 7 – *Visualisation pour Stéphane Mallarmé générée par la vue " : :Logilab : :Person"*

bussy sont affichées et non plus l'ensemble des triplets, comme c'est le cas pour la visualisation proposée par DBPedia. De plus, ici, un certain nombre de ressources liées sont présentées dans cette visualisation mais les ressources liées ne sont pas récupérées directement lors de la négociation de contenu. Elles sont interrogées, par une requête AJAX, dès l'affichage de la visualisation de la vue. C'est pour cela que nous pouvons ici observer des URI affichées à côté des relations vers des ressources liées. Par exemple dans la catégorie "Influenced by" l'URI correspondante à la ressource de Stéphane Mallarmé est affichée. Dès que les données correspondantes à cette ressource sont récupérées, le nom de la personne est présentée à la place de l'URI. Le même fonctionnement est appliqué pour chaque URI présentée, mais aussi pour la photo de Claude Debussy. La figure 6 présente la même visualisation mais avec toutes les ressources liées récupérées.

Dans cette visualisation, nous pouvons voir que Stéphane Mallarmé a influencé Claude Debussy. La chaîne de caractères "Stéphane Mallarmé" a été récupérée en interrogeant la ressource référencée par l'URI "`http://dbpedia.org/resource/Stéphane_Mallarmé`". Cette chaîne de caractères est affichée comme un lien pointant vers la ressource associée. Il est donc possible de cliquer directement sur ce lien pour accéder à la visualisation de la ressource pour Stéphane Mallarmé. Stéphane Mallarmé étant aussi une personne, la visualisation utilise la même vue. Cette visualisation est présentée sur la figure 7. Nous avons vu qu'il est possible de passer d'une visualisation d'un document HTML vers une ressource RDF. Il est aussi possible de passer d'une ressource RDF vers une autre ressource RDF. Tout le traitement de sélection de vue et de génération de la visualisation étant effectué côté client, tout est dynamique. Par conséquent, il est possible d'effectuer le chemin inverse, c'est à dire d'une ressource RDF vers un document HTML.

4.2 Visualisation des données de la conférence SemWeb.pro

Une version spécifique du site de la conférence SemWeb.pro 2018 [16] a été déployée afin de permettre la négociation de contenu sur toutes les pages. Des vues spécifiques ont également été développées et rendues disponibles sur le serveur de vues "`https://views.semweb.pro/`". Le même procédé que pour la page de Claude Debussy est appliqué, mais cette fois ci sur le document HTML répertoriant les communications lors de SemWeb.pro 2018. La figure 8 présente la visualisation HTML générique du site de la conférence générée par le serveur. En activant l'extension la vue " : :Logilab : :Conference" est utilisée pour générer la visualisation présentée sur la figure 9.

Si nous cliquons par exemple sur la présentation se nommant "Un navigateur pour le web des données" nous arrivons sur une visualisation générée par la vue " : :Logilab : :Conference-

16. `http://demo.semweb.pro`

IC 2018

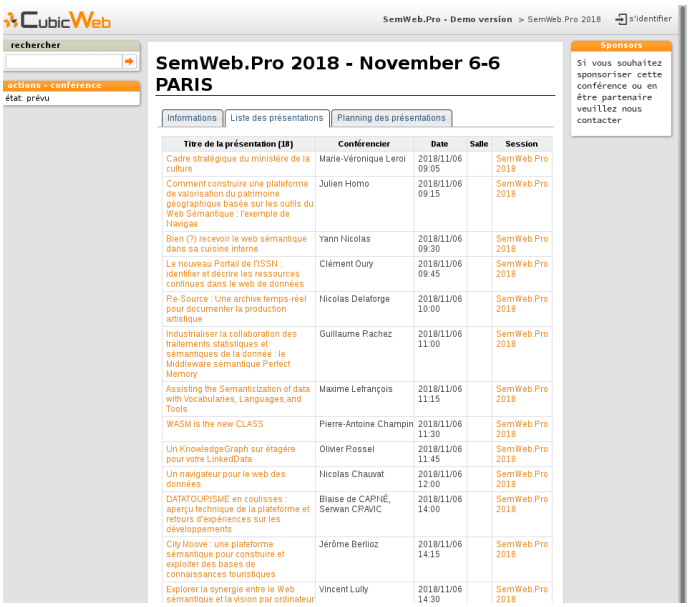


FIGURE *8 – Visualisation générée par le serveur pour la conférence SemWeb.pro 2018*

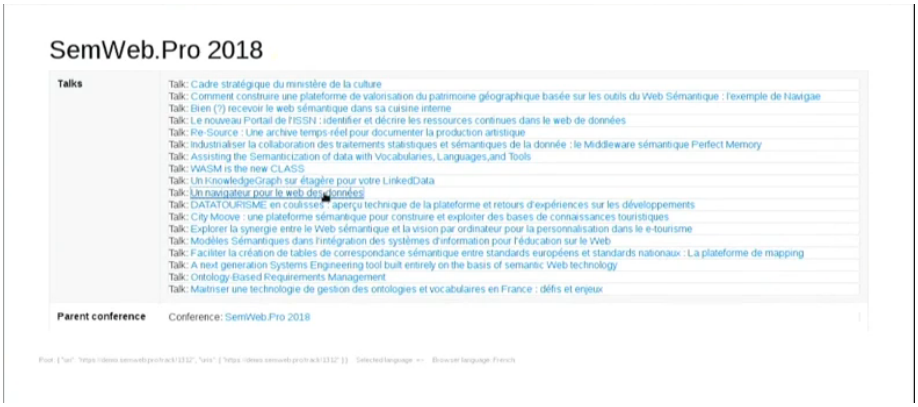


FIGURE *9 – Visualisation générée par la vue " : :Logilab : :Conference" pour la conférence SemWeb.pro 2018*

FIGURE *10 – Visualisation d'une communication de la conférence SemWeb.pro 2018, générée par la vue " : :Logilab : :ConferenceTalk"*

FIGURE *11 – Visualisation d'un auteur d'une communication de la conférence SemWeb.pro 2018 en utilisant la vue " : :Logilab : :Person"*

Talk". Cette visualisation est présentée sur la figure 10. Nous observons que le changement de ressource dans l'extension peut aussi amener à changer de vue pour générer la visualisation. Si nous cliquons sur l'auteur de la communication ("Nicolas Chauvat") alors nous pouvons observer une visualisation générée pour des ressources de type "$foaf : Person$". La vue utilisée sera donc " : :Logilab : :Person", comme pour la ressource "Claude_Debussy" provenant de DBPedia. Une vue créée pour cette extension peut donc être utilisée pour un type de ressource en particulier, quel que soit le serveur qui distribue cette ressource. La visualisation pour "Nicolas Chauvat" est présentée sur la figure 11.

Sur cette figure il est possible de voir que la conférence SemWeb.pro ne détient que très peu d'informations sur Nicolas Chauvat. Mais aussi que Nicolas Chauvat semble connaître Claude Debussy. Cette ressource est celle provenant de DBPedia. Si nous cliquons sur ce lien nous tomberons sur la figure 6. Cet exemple illustre le passage transparent d'un serveur à un autre en suivant les liens affichés et en réutilisant la même vue pour mettre en page des données de même type indépendamment du serveur dont elles proviennent. Il illustre aussi le fait que les mêmes données, ici la description d'une conférence, peuvent être mises en page soit par le serveur (cf figure 8), soit par le navigateur (cf figure 9).

5 Conclusion et perspectives

5.1 Un navigateur pour le web des données liées

Le web est un système décentralisé d'identification (URL) et d'échange (HTTP) d'objets numériques qui peuvent être des documents (HTML) ou des données (RDF). Nous avons cherché à disposer d'un navigateur avec lequel parcourir de façon homogène et transparente l'ensemble du Web, en suivant à chaque fois les liens qui pointent d'une ressource à la suivante, qu'elle soit constituée par un document HTML ou des données RDF. Nous avons développé une WebExtension qui permet, lorsque l'URL de la barre d'adresse désigne des données RDF, de mettre en page automatiquement les données reçues pour les visualiser. Pour cela, nous avons défini la notion de "vue" comme étant une fonction javascript qui prend en entrée un graphe RDF et produit en sortie un document HTML. Une fois installée, l'extension du navigateur est configurée en y ajoutant ces vues publiées statiquement sur des serveurs HTTP. Lors du chargement de données RDF, un mécanisme de sélection choisit la vue la plus adaptée et l'applique pour que les données reçues soient mises en page et que le HTML résultant soit affiché. Ceci a lieu côté navigateur, sans interaction avec le serveur. Nous avons ainsi obtenu un navigateur aux fonctionnalités étendues, avec lequel la séparation entre la publication des ressources par le serveur et leur visualisation par le client est plus

nette qu'elle ne l'était jusqu'à présent, ce qui présente plusieurs avantages.

Le premier avantage est la possibilité de naviguer de façon plus homogène parmi des sites qui utilisent les mêmes ontologies. Comme les vues ne dépendent pas du serveur depuis lequel les données sont téléchargées, le sentiment de parcourir un ensemble cohérent d'informations est nettement renforcé, même si l'on passe d'un serveur à un autre durant la navigation, car les données de même nature donnent lieu au même affichage. Ceci nous rapproche de la consultation d'un graphe de données global plutôt que d'un ensemble de silos interconnectés.

Le deuxième avantage est la possibilité de personnaliser les interfaces homme-machine, puisque la sélection de la visualisation est faite dans le navigateur à partir des vues configurées. Deux utilisateurs peuvent donc faire des choix de configuration différents et ne pas avoir le même affichage quand ils consultent la même URL qui désigne des données RDF. Chacun pourra donc disposer d'une vue adaptée à ses préférences ou à sa tâche en cours.

Le troisième avantage est d'abaisser globalement le coût de l'échange d'information entre serveur et client, puisqu'il devient possible de mutualiser le coût de la mise au point de la visualisation. Là où chaque organisme souhaitant partager des données avait l'obligation de développer et de maintenir une interface utilisateur, nous proposons des vues qui ne dépendent que du type des données et peuvent être utilisées avec un nombre quelconque de serveurs. Un organisme qui souhaite partager ses données peut donc le faire à moindre coût en choisissant des ontologies pour lesquelles plusieurs des vues nécessaires sont déjà disponibles et utilisées.

5.2 Perspectives

Bien que l'extension décrite ci-dessus soit d'ores et déjà utilisable, nous avons identifié plusieurs pistes d'amélioration.

Diagnostic en cas d'erreur

Notre extension est fonctionnelle, mais dans de nombreux cas, ne donne pas les résultats escomptés. La forme RDF des données n'est pas trouvée, la vue espérée n'est pas sélectionnable, les liens ne désignent pas les ressources attendues, etc. Les raisons peuvent être multiples : absence des en-têtes adéquates, erreur dans la négociation de contenu, mécanisme de redirection mal mis en oeuvre, triplets RDF manquants, etc. Pour établir un diagnostic facilement, il est nécessaire que l'extension fournisse un maximum d'informations sur les échanges qui ont eu lieu avec le serveur HTTP. Nous avons commencé à développer une barre latérale qui peut afficher ces informations et poursuivrons ce travail rapidement.

Simplification du développement et du déploiement des vues

Le développement et le déploiement de nouvelles vues doit être aussi simple que possible. Le choix qui a été fait d'utiliser des fonctions javascript pour les vues laisse une grande latitude au développeur et n'impose pas de changement radical par rapport aux méthodes et outils ayant cours actuellement. Fournir des exemples d'utilisation des librairies les plus répandues pourrait faciliter l'adoption. Une documentation existe déjà, mais elle est améliorable et pourrait être assortie de plus d'exemples et de modèles à partir desquels démarrer un nouvel ensemble de vues.

Compatibilité avec Fresnel

Comme décrit dans la section 3.2.3, la sélection de vue se fait actuellement avec un système de priorité. Lorsque l'extension est configurée avec des vues issues de plusieurs serveurs, il n'existe aucune garantie quant à la cohérence des priorités fournies par ces différentes vues. Donner un score de priorité à une vue sans connaître ni les critères ni les autres valeurs n'est pas chose aisée et peut amener à des conflits et des sélections erronées. Bien que l'utilisateur puisse changer de vue s'il le souhaite, il serait intéressant d'améliorer ce système

de sélection de vue pour toujours avoir la vue la plus pertinente sans intervention de l'utilisateur. Une solution pourrait être d'associer une expression Fresnel Selector Language (FSL) à chaque vue et d'utiliser ces expressions au moment de sélectionner la meilleure vue possible pour les données reçues.

Algorithme de selection de vue

Il est possible aussi de trouver une meilleure méthode pour la selection de la vue la plus pertinente. Une idée serait de pouvoir proposer à l'utilisateur du navigateur de pouvoir ordonner les serveurs de vues pour définir une priorité. L'extension essaierait d'abord de trouver une vue pertinente dans le premier avant d'étudier le deuxième, et ainsi de suite. Les scores de priorité des vues seraient uniquement en fonction des autres vues du même serveur et non de toutes les vues ajoutées par l'utilisateur.

Participation à la standardisation de la navigation RDF

Nous avons vu dans la section 3.1.1 que nous utilisons soit un paramètre de la réponse HTTP soit une balise dans le code HTML pour vérifier la disponibilité de la ressource dans un autre format. Il serait intéressant de couvrir toutes les possibilités de référence à des formats alternatifs pour la négociation de contenu et surtout de pouvoir respecter une méthode standard, si elle existe ou si une convention finit par apparaître.

API Hypermedia

Nous nous sommes concentrés sur la lecture des ressources publiées sur le Web, mais sommes bien conscients de l'importance de l'écriture et de la modification des données. Nous comptons explorer les relations entre nos travaux et les standards existants pour les API Hypermedia et la manipulation de RDF, parmi lesquels HAL, Hydra et Linked Data Platform.

Vues pour la consultation d'entrepôts RDF

Comme expliqué dans la section 2, nous avons distingué la navigation au sein du Web de la consultation d'un graphe RDF publié par exemple via un point d'accès SPARQL. Nous pensons que les vues telles que nous les avons définies et employées pourraient être réutilisées par un outil de consultation générique de graphe RDF.

Industrialisation

Nous souhaitons aussi obtenir un système suffisamment efficace, fiable et robuste pour qu'il puisse être utilisé au quotidien par nos clients. Nous développons aujourd'hui principalement en utilisant CubicWeb, qui propose un système de vues générées par le serveur. Notre objectif serait de les supprimer au profit de vues génériques ou spécifiques ajoutées au navigateur comme décrit ici. En découplant la publication des données du déploiement des vues, nous espérons faciliter leur modification et donc accroître notre capacité à livrer des changements en continu, comme encouragent à le faire les méthodes agiles que nous pratiquons.

Collaborations et logiciel libre

Le code source que nous avons développé est publié sous la licence LGPLv3 à l'url `https://www.cubicweb.org/project/cubicweb-linked-data-browser`. Nous sommes ouverts à toutes formes de collaboration.

IC 2018

Remerciements

Nous souhaitons remercier Noé Gaumont et Frank Bessou, qui sont nos collègues chez Logilab et qui ont participé à la rédaction, à la structuration et à la relecture de cet article. Nous remercions aussi Olivier Cayrol qui a su trouver le budget nécessaire à la rédaction de cet article. Enfin nous remercions tous les contributeurs de CubicWeb pour leurs idées et en particulier Sylvain Thénault, qui en a été le développeur principal pendant de longues années.

Références

BERNERS-LEE T., HOLLENBACH J., LU K., PRESBREY J. & SCHRAEFEL M. (2008). Tabulator redux : Browsing and writing linked data. *loutre*.

CORBY O. & ZUCKER C. F. (2015). Sttl : A sparql-based transformation language for rdf. In *11th International Conference on Web Information Systems and Technologies*.

DESTANDAU M. (2016). Vizskos, a visualizer for skos based thesaurus. *In Cartographie meetup, Paris*.

HUYNH D., MAZZOCCHI S. & KARGER D. (2007). Piggy bank : Experience the semantic web inside your web browser. *Web Semantics : Science, Services and Agents on the World Wide Web*, **5**(1), 16–27.

LOHMANN S., NEGRU S., HAAG F. & ERTL T. (2016). Visualizing ontologies with vowl. *Semantic Web*, **7**(4), 399–419.

MANSOUR E., SAMBRA A. V., HAWKE S., ZEREBA M., CAPADISLI S., GHANEM A., ABOULNAGA A. & BERNERS-LEE T. (2016). A demonstration of the solid platform for social web applications. In *Proceedings of the 25th International Conference Companion on World Wide Web*, p. 223–226 : International World Wide Web Conferences Steering Committee.

PIETRIGA E. (2003). Isaviz : A visual authoring tool for rdf. *World Wide Web Consortium.[Online]. Available : http ://www. w3. org/2001/11/IsaViz*.

PIETRIGA E., BIZER C., KARGER D. & LEE R. (2006). Fresnel : A Browser-Independent Presentation Vocabulary for RDF. In D. HUTCHISON, T. KANADE, J. KITTLER, J. M. KLEINBERG, F. MATTERN, J. C. MITCHELL, M. NAOR, O. NIERSTRASZ, C. PANDU RANGAN, B. STEFFEN, M. SUDAN, D. TERZOPOULOS, D. TYGAR, M. Y. VARDI, G. WEIKUM, I. CRUZ, S. DECKER, D. ALLEMANG, C. PREIST, D. SCHWABE, P. MIKA, M. USCHOLD & L. M. AROYO, Eds., *The Semantic Web - ISWC 2006*, volume 4273, p. 158–171. Berlin, Heidelberg : Springer Berlin Heidelberg.

QUAN D., HUYNH D. & KARGER D. R. (2003). Haystack : A platform for authoring end user semantic web applications. In *International Semantic Web Conference*, p. 738–753 : Springer.

SIMON A., WENZ R., MICHEL V. & DI MASCIO A. (2013). Publishing bibliographic records on the web of data : opportunities for the bnf (french national library). In *Extended Semantic Web Conference*, p. 563–577 : Springer.

Une Ontologie des Processus Métier (BBO)

Amina Annane, Nathalie Aussenac-Gilles, Mouna Kamel

IRIT, CNRS – Université de Toulouse, Toulouse, France
prénom.nom@irit.fr

Résumé : Les activités de recherche dans le domaine de la Modélisation de Processus Métier ont donné lieu à différentes propositions de modèles et ontologies, certains étant reconnus comme standards. Mais aucun de ces modèles n'est, à notre connaissance, suffisamment riche pour permettre à un agent virtuel de superviser l'exécution d'un processus métier et d'exploiter des Retours d'EXpériences (REX) en cas d'anomalie. Notre participation au projet AVI-REX, en partenariat avec de grandes entreprises industrielles, a pour objectif de construire une telle ontologie. Le travail présenté dans cet article correspond à la première étape de la construction, à savoir la description de processus métier, avec toutes les connaissances requises pour leur exécution. Pour répondre aux besoins spécifiés, nous nous sommes basés sur plusieurs modèles existants, que nous avons intégrés puis enrichis. L'ontologie obtenue, appelée BBO, a été évaluée selon différents critères, puis mise à disposition de la communauté.

Mots-clés : Modélisation de processus métier, Web sémantique, Ontologie, Intégration de données

1 Introduction

Les processus métier représentent le savoir-faire des entreprises : "*a process is a particular procedure for doing something involving one or more steps or operations. The process may produce a product, a property of a product, or an aspect of a product*" (ISO 10303-49). La Modélisation de Processus Métier (BPM pour Business Process Modeling en anglais) permet d'analyser, améliorer, simuler et automatiser ces processus (Rospocher *et al.*, 2014). A l'heure de l'industrie 4.0 où l'ambition est de rendre la communication toujours plus efficace et performante, d'une part entre les différents systèmes informatiques, et d'autre part, entre les humains et ces systèmes, bien modéliser les processus prend de plus en plus d'importance.

Les technologies du web sémantique constituent des solutions prometteuses pour réaliser cette ambition (Vogel-Heuser & Hess, 2016). Les ontologies notamment permettent de décrire les connaissances de façon formelle (Berners-Lee *et al.*, 2001) et possèdent des capacités de raisonnement pouvant assurer la cohérence des processus métiers (Rospocher *et al.*, 2014; Roy *et al.*, 2018). De plus, représenter les processus métiers à l'aide d'ontologies semble bien répondre à la question de la communication et de l'interopérabilité entre différents systèmes.

Dans ce contexte, le projet AVI-REX vise la réalisation d'un modèle d'agent virtuel pour superviser et guider l'exécution des processus métier, s'inscrivant ainsi dans l'ère de l'industrie 4.0. Ce projet regroupe l'entreprise SimSoft Industry, qui conçoit et implémente l'agent virtuel, l'IRIT [1] et deux entreprises industrielles futures utilisatrices de ces agents : Thales Alenia Space (TAS) et Continental. Le rôle prévu de cet agent virtuel sera le suivant :

1. guider pas à pas l'opérateur et contrôler le bon déroulement de l'exécution des processus métiers ;
2. répondre aux questions que l'opérateur peut poser sur l'exécution du processus ;
3. exploiter les REX (Retours d'EXpérience) et aider les experts métiers à poser un diagnostic en cas d'anomalie survenue lors de l'exécution des processus.

Pour une entreprise donnée, le modèle d'agent est adapté au contexte du poste de travail dans lequel il doit s'intégrer, et cela en lui fournissant des connaissances (sous la forme d'une base de connaissances) propres aux procédures réalisées à ce poste et aux ressources utilisées. Notre contribution au sein d'AVI-REX consiste à construire une ontologie permettant de décrire le plus finement possible les processus métier, ainsi que les traces de leurs

1. https ://www.irit.fr

exécutions pour réaliser la base de connaissances d'un tel agent virtuel. C'est cette ontologie qui sera alors livrée aux clients, qui auront à charge de la spécialiser et de la peupler selon les caractéristiques de leurs processus métiers respectifs.

Dans cet article, nous détaillons le processus qui, après un état de l'art et une spécification des besoins, nous a conduit à construire une ontologie de domaine. Nous présentons ici uniquement la partie de l'ontologie décrivant des processus métier, en détaillant les connaissances nécessaires qui permettront, plus tard, de modéliser les retours d'expériences. Dans la suite, nous désignerons cette partie d'ontologie simplement par "l'ontologie" ou "BBO". La construction de BBO s'appuie sur les quatre étapes majeures (Figure 1) qu'intègrent la plupart des méthodes de construction d'ontologies (Fernández *et al.*, 1997) :

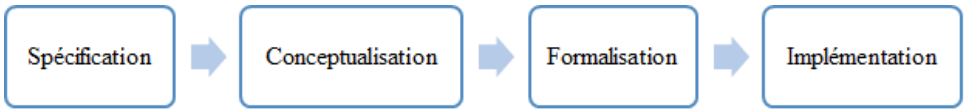


FIGURE *1 – Principales étapes de la construction d'ontologie*

Suivant ce canevas, nous avons :
— spécifié les besoins que doit satisfaire l'ontologie à l'aide de documents techniques et d'experts du domaine, ainsi que d'un ensemble de questions portant sur les compétences techniques (ou "competency questions") ;
— identifié les modèles et ressources existants et réutilisables, puis spécialisé certains concepts de ces modèles ;
— formalisé le contenu de ces modèles ainsi qu'un ensemble de spécifications décrites en Langage Naturel (LN) associées aux modèles réutilisés, ou ayant été exprimées par les experts du domaine ;
— implémenté, documenté et évalué l'ontologie.

Dans la suite de l'article, nous dressons un état de l'art (section 2), puis nous décrivons chacune des étapes de la construction de BBO : la section 3 présente les spécifications, la section 4 la conceptualisation, la section 5 la formalisation et enfin la section 6 l'évaluation. Nous concluons notre article et donnons les suites envisagées pour ce travail en section 7.

2 Travaux connexes

Nous avons alors cherché dans l'état de l'art des ontologies et modèles à intégrer ou aligner pour construire une ontologie des processus métier qui soit à la fois générale (pour toute entreprise) et très précise dans la description des ressources, agents et équipements utilisés.

Le modèle qui semble faire référence dans le domaine est BPMN (OMG, 2011), qui bénéficie de spécifications assez précises et d'une logique d'exécution bien définie. Une partie de ces spécifications est exprimée en UML et au format XML, et une partie en langage naturel. Les spécifications de BPMN 1.0 (datant de 2008) ont donné lieu à la construction semi-automatique d'une ontologie BPMN 1.0, qui a été enrichie par des axiomes et documentée manuellement (Rospocher *et al.*, 2014). Les nouvelles spécifications de 2011, qui intègrent cette fois la sémantique relative à l'exécution des processus, ont également donnée lieu à la construction d'ontologies, comme l'ontologie BPMN 2.0 présentée dans (Natschläger, 2011) qui, à notre connaissance, n'est pas accessible à la communauté. Une autre ontologie a été extraite des spécifications BPMN 2.0 [2], mais sans prise en compte du texte décrivant les classes ni documentation. Pour tous ces travaux, l'idée était de traduire tout le méta-modèle BPMN en une ontologie, ce qui n'est pas notre objectif, certaines parties étant inutiles au regard de nos besoins. En revanche, plusieurs auteurs soulignent que le modèle BPMN ne couvre pas

2. https ://dkm.fbk.eu/bpmn-ontology

toutes les connaissances requises pour représenter des processus métier : il manque des éléments pour décrire des ressources, les sites de production ou les agents réalisant les processus (Awad *et al.*, 2009; Stroppi *et al.*, 2011; Marcinkowski & Kuciapski, 2012; Bocciarelli *et al.*, 2016). Pour cela, nous avons identifié d'autres ontologies pertinentes : une ontologie pour la gestion des projets logiciels (Ruíz *et al.*, 2004), une ontologie des processus logiciels (Falbo & Bertollo, 2009), une ontologie des processus de maintenance industrielle (Karray *et al.*, 2012), une ontologie des processus de fabrication (Chungoora *et al.*, 2013). Enfin, les descriptions des processus métier de certaines ontologies (Uschold *et al.*, 1998; Pedrinaci *et al.*, 2008; Cabral *et al.*, 2009) sont déjà couvertes par le modèle BPMN, avec un grain plus fin.

Nous avons donc retenu BPMN 2.0 comme modèle de référence, et identifié des ontologies pour en combler certaines lacunes. Nous avons décidé de construire l'ontologie BBO à partir de ces sources faute d'ontologie de référence pour représenter précisément les processus métier et les retour d'expériences.

3 Spécifications

Rappelons que BBO devra pouvoir être exploitée tant pour guider un opérateur lors de l'exécution pas à pas d'un processus métier, que pour aider les experts à poser un diagnostic en cas d'anomalie ou de panne lors de l'exécution d'un processus métier. Afin de bien cerner les exigences à satisfaire, nous avons analysé différentes sources de connaissances.

3.1 Documents techniques

Les deux partenaires industriels, TAS et Continental, ont fourni chacun deux corpus :
- un corpus *PM_TAS* et un corpus *PM_Continental* décrivant des processus métier, soit 20 processus métier décrits au total,
- un corpus *REX_TAS* et un corpus *REX_Continental* décrivant des retours d'expérience (REX), soit 28 REX au total.

L'extrait d'un processus métier de chez TAS (que nous avons rendu anonyme) de la table 1 montre un processus composé de 3 activités à exécuter séquentiellement. Une activité peut, à son tour, être composée d'une ou plusieurs actions (e.g., activité 3). Nous appelons *tâche* ou *action* l'élément atomique dans la description d'un processus. Une action (caractérisée par un verbe) peut faire référence à différents types de ressources (*fichier*, *logiciel*, *dispositif*, etc.), à des valeurs de paramètres (*code erreur*), à des instructions conditionnelles (*si l'exécution a généré...*), etc. Le lieu d'exécution du processus est également indiqué (*sur la station MMA*). On note également un cas de polysémie : la station MMA représente à la fois une ressource et un lieu (i.e., un poste de travail). Ce petit exemple montre la finesse de description requise. Ces documents techniques ont été analysés avec l'aide des experts impliqués dans le projet.

```
Sur la station MMA :
1. Exécuter le fichier MUXF.exe
2. En utilisant le logiciel SOFT, vérifier si l'exécution
du script a généré une alarme.
3. Si l'exécution a généré une alarme, exécuter le
processus de résolution d'anomalie N°XY
```

TABLE *1 – Extrait anonyme d'un processus métier TAS*

3.2 Questions basées sur les compétences techniques

Les "competency questions" sont reconnues pour être un bon moyen de spécifier les besoins d'une ontologie (Grüninger & Fox, 1995). Un ensemble de questions de ce type a

émergé suite à notre collaboration avec les experts du projet. Cet ensemble a été enrichi de questions de même type issues de la littérature (Falbo & Bertollo, 2009; Abdalla *et al.*, 2014). Nous avons obtenu 20 questions principales [3], dont nous donnons un extrait dans la Table 2. Ces questions ont permis de mettre en évidence l'importance de certaines entités comme les activités, les ressources, etc.

Quelles ressources sont nécessaires à une activité ?
Quelles ressources sont produites par une activité ?
En quelles sous-activités une activité peut-elle se décomposer ?
Quelle est la nature/le type d'une ressource ?
Quelles activités doit précéder/suivre une activité donnée ?
Qui doit exécuter une activité donnée ?
Où l'activité doit-elle être exécutée ?
...

TABLE *2 – Extrait de l'ensemble des questions basées sur les compétences techniques*

3.3 Synthèse

L'étude de ces sources de connaissance a permis de circonscrire les exigences auxquelles devait répondre l'ontologie, et de caractériser la notion de processus métier. Un processus métier se décompose en un ensemble d'activités (en moyenne une vingtaine), une activité pouvant être un sous-processus ou une tâche ; la tâche est l'élément atomique réalisé par un agent (i.e., opérateur). Une tâche peut nécessiter ou produire une ou plusieurs ressources de natures diverses : matérielles, logicielles, humaines, données, etc. Les activités doivent pouvoir s'exécuter en séquence, en parallèle ou de façon itérative. Elles doivent pouvoir être contrôlées par des évènements (interruption, reprise, etc.). Il est important de connaître la composition et le site de production (entreprise, station de travail, etc.) des produits manufacturés, ainsi que la valeur de certains paramètres qui peuvent conditionner l'exécution des tâches. Enfin, il est important de spécifier le rôle (responsabilité ou autre) de l'agent qui peut/doit réaliser une activité donnée. L'agent peut être une ressource humaine ou une ressource logicielle.

4 Conceptualisation

Suite à l'analyse des spécifications, nous avons identifié cinq entités principales, représentées par des concepts de BBO : *Processus*, *Entrées/Sorties des activités*, *Agent*, *Produit Manufacturé* et *Site de production*. Dans la suite, nous présentons les modèles conceptuels associés à chacun de ces concepts en utilisant des diagrammes de classes UML.

Construire une ontologie "from scratch" est coûteux en termes de temps et d'effort. L'état de l'art présenté en section 2 nous a permis d'identifier les ressources pertinentes pour répondre aux besoins et que nous avons réutilisées : (i) le modèle BPMN 2.0 (OMG, 2011), standard dans le domaine de la représentation des processus métier, qui constituera le noyau de l'ontologie ; (ii) la taxonomie des ressources proposée par (Falbo & Bertollo, 2009; Karray *et al.*, 2012) ; (iii) la taxonomie des sites de production de (Chungoora *et al.*, 2013) et (Fraga *et al.*, 2018) ; (iv) le fragment d'ontologie des agents présent dans (Ruíz *et al.*, 2004).

Pour plus de lisibilité, nous avons scindé la présentation de BBO en cinq sous-sections, chacune correspondant à un concept pivot cité ci-dessus. Par ailleurs, pour différencier les concepts issus du modèle BPMN (modèle ayant servi de base à BBO) des autres sur les diagrammes de classes, nous avons (i) souligné les concepts issus d'ontologies existantes (cf. le concept `UO:unit` de la Figure 4) et (ii) encadré les nouveaux concepts que nous avons ajoutés (cf. le concept `Parameter` de la Figure 4).

3. Liste des 20 questions : `https://github.com/AminaANNANE/BBO_BPMNbasedOntology`

4.1 Processus

Processus est le concept principal de BBO. Le modèle qui nous a semblé le plus abouti pour représenter les processus métier est le modèle BPMN 2.0 (OMG, 2011). En effet, en plus de son expressivité pour représenter un processus, ce modèle intègre la sémantique d'exécution de processus. Il nous a donc paru intéressant de réutiliser une partie de ce modèle, en laissant de côté les concepts liés à la représentation graphique et aux interactions entre processus (i.e., collaboration, conversation, choreography, etc.). Les principaux concepts et relations du fragment réutilisé sont décrits en Figure 2.

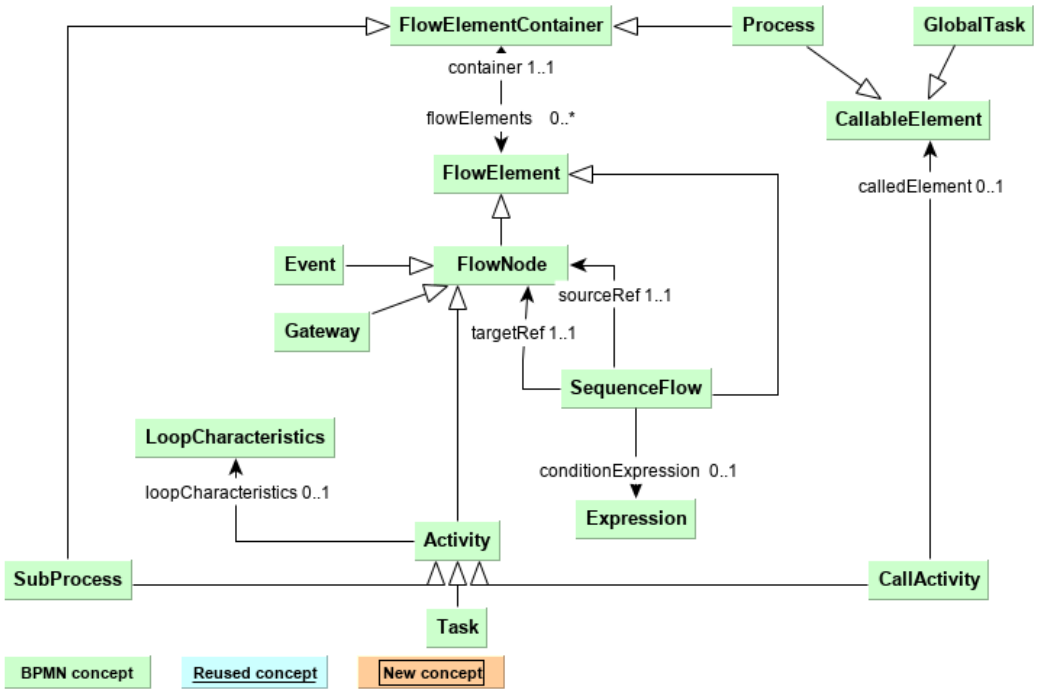


FIGURE 2 – *Diagramme de classes correspondant à la représentation des processus*

Un processus (*Process*) est considéré comme un regroupement (*FlowElementContainer*) d'éléments de flux (*FlowElement*) de deux types :

— *FlowNode* qui désigne une activité (*Activity*), un évènement (*Event*), une passerelle (*Gateway*)
— *SequenceFlow* qui correspond à un connecteur entre deux *FlowNode*. Un connecteur peut être soumis à condition (*Expression*).

Une activité (*Activity*) représente le travail à faire. Elle est spécialisée en trois sous-classes :

— une tâche atomique (*Task*)
— une tâche complexe regroupant plusieurs tâches atomiques (*SubProcess*)
— Une activité d'appel (*CallActivity*) qui appelle un autre processus (*Process*) ou une tâche réutilisable (*GlobalTask*) : les sous classes de *CallableElement*.

La classe *LoopCharacteristics* représente une activité à exécuter de façon itérative.

Un évènement (*Event*) peut survenir au cours de l'exécution d'un processus : le processus peut par exemple être interrompu ou reprendre son exécution s'il avait été interrompu. Un évènement a généralement une cause ou un impact. Différents types d'évènements peuvent être spécifiés : les évènements temporels (*timerEvent*), les évènements conditionnels (*conditionalEvent*), etc.

Une passerelle (*Gateway*) sert à contrôler comment les connecteurs (*SequenceFlow*) interagissent, par exemple s'ils convergent ou s'ils divergent (Figure 3).

Par ailleurs, nous avons enrichi ce modèle selon les spécifications de BPMN 2.0 exprimées en langage naturel. En effet, dans ces spécifications, BPMN décrit les propriétés de certains concepts à l'aide d'attributs valués. Par exemple, l'entité *Gateway* (passerelle) possède un attribut *GatewayDirection* dont la valeur ("convergent", "divergent", "mixte", ou "inconnu") détermine le type. Chaque type de passerelle a alors ses propriétés avec des contraintes propres : les passerelles convergentes doivent être la cible d'au moins deux instances de connecteurs (*SequenceFlow*) et la source d'un seul connecteur, les passerelles divergentes doivent être la cible d'une seule instance de connecteur et la source d'au moins deux connecteurs, etc. (Figure 3). La simple affectation d'une valeur à un attribut n'assure pas la cohérence avec la définition. Pour ces cas-là, nous avons appliqué la règle qui consiste à créer, pour tout attribut *att* de la classe *C*, autant de sous-classes de *C* que de valeurs possibles pour *att*. Pour le concept *Gateway* par exemple, nous avons créé 4 sous-classes correspondant aux quatre valeurs d'attribut.

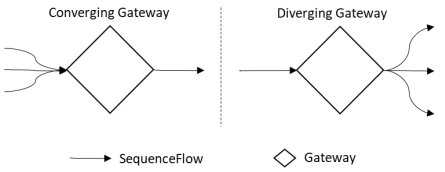


FIGURE *3 – Passerelle convergente vs. passerelle divergente*

Ce même principe a été appliqué à d'autres concepts présentant les mêmes caractéristiques que *Gateway*. La table 3 montre quelques exemples d'enrichissements effectués à partir des spécifications BPMN en langage naturel.

Classe	Sous-Classes
SequenceFlow	ConditionalSequenceFlow, DefaultSequenceFlow, NormalSequenceFlow
SubProcess	EventBasedSubProcess
Gateway	ConvergingGateway, DivergingGateway, MixedGateway, UnspecifiedGateway
EventBasedGateway	ExclusiveEventBasedGateway, ParallelEventBasedGateway
Event	CancelEvent, CompensationEvent, ConditionalEvent, ErrorEvent, EscalationEvent, LinkEvent, MultipleEvent, NoneEvent, ParallelMultipleEvent, SignalEvent, TerminateEvent, TimerEvent

TABLE *3 – Spécialisation des classes BPMN*

Bien que BPMN 2.0 permette de décrire de façon tout à fait satisfaisante les processus métiers, il ne permet pas de représenter certaines connaissances comme les différents types de ressources, les produits manufacturés et les sites des activités, qui, selon les spécifications, sont nécessaires. Par exemple, il n'est pas possible de représenter le fait que la tâche (2) de l'exemple Figure 1 nécessite la ressource "SOFT" de type Logiciel (Software), de donner une valeur attendue à un paramètre d'une ressource donnée, ou encore d'exprimer que les trois tâches doivent être exécutées sur la "station MMA". Il est donc nécessaire d'enrichir ce fragment d'ontologie avec de nouveaux concepts et de nouvelles relations.

4.2 Les Entrées/Sorties des activités

Une activité nécessite des entrées et produit des sorties, sachant que ces entrées/sorties peuvent être de types différents et peuvent être caractérisées par plusieurs valeurs de paramètres. Or le modèle BPMN ne permet de spécifier que des données comme entrées ou sorties d'une activité (*DataInput* et *DataOutput*). Nous avons donc rajouté deux relations "has_resourceInput" et "has_resourceOutput" entre les concepts *InputOutputSpecification* et *Resource* (voir Figure 4).

Par ailleurs, pour représenter les paramètres, leurs valeurs et leurs unités de mesure, nous avons rajouté le fragment d'ontologie décrit dans la Figure 4 (les concepts avec des labels encadrés). Les concepts *Unit* et *Prefix* proviennent de l'ontologie des unités de mesure (UO) [4].

4. http ://purl.obolibrary.org/obo/uo.owl, visited in 2019/03

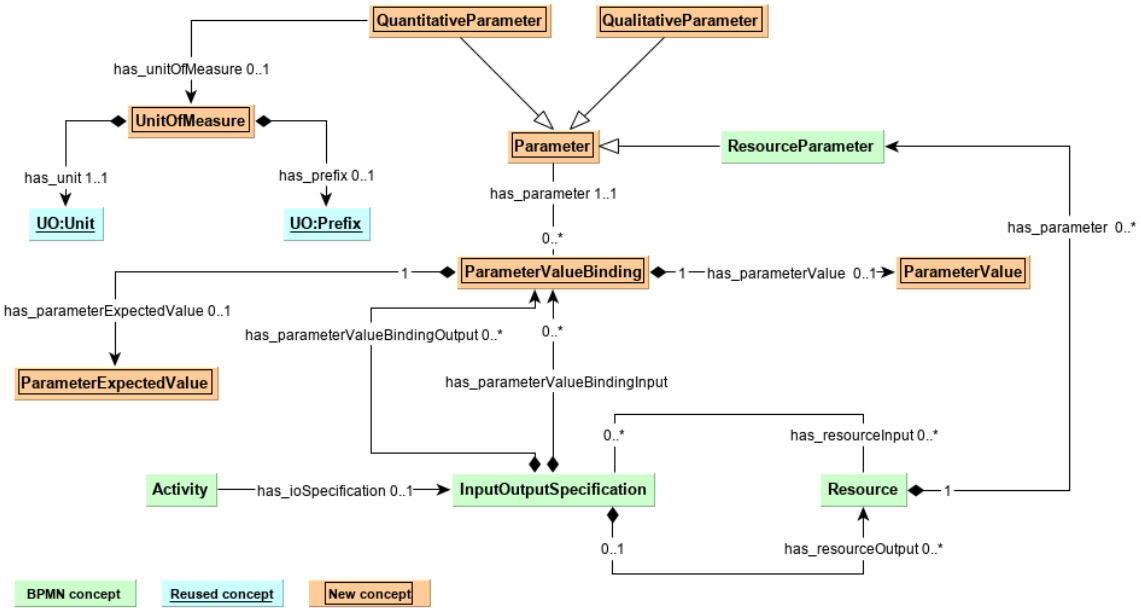


FIGURE *4 – Diagramme de classes correspondant aux entrées/sorties d'activité*

Le concept *Resource* existe dans les spécifications du modèle BPMN 2.0, mais sa sémantique est ambiguë. En effet, la classe *Resource* est supposée couvrir tout type de ressource (p. 95) : "*The Resource class is used to specify resources that can be referenced by Activities. These Resources can be Human Resources as well as any other resource assigned to Activities during Process execution time.*" Mais les relations qui lient les ressources aux processus (p. 148) ou aux activités (p.152), limitent les ressources aux agents responsables de l'exécution du processus : "*... defines the resource that will perform or will be responsible for the Process. The resource, e.g., a performer, can be specified in the form of a specific individual, a group, an organization role or position, or an organization.*" Cette définition semble être la plus adoptée. En effet, dans la plupart des travaux qui font référence à BPMN, la notion de ressource est équivalente à celle d'agent (Awad *et al.*, 2009; Stroppi *et al.*, 2011).

Dans BBO, nous adoptons la première définition du concept *Resource*, qui englobe tout type de ressources afin de spécifier les différents types d'entrées/sorties. En effet, nous inspirant des travaux de (Falbo & Bertollo, 2009; Karray *et al.*, 2012), nous avons défini une taxonomie des ressources que nous présentons dans la Figure 5.

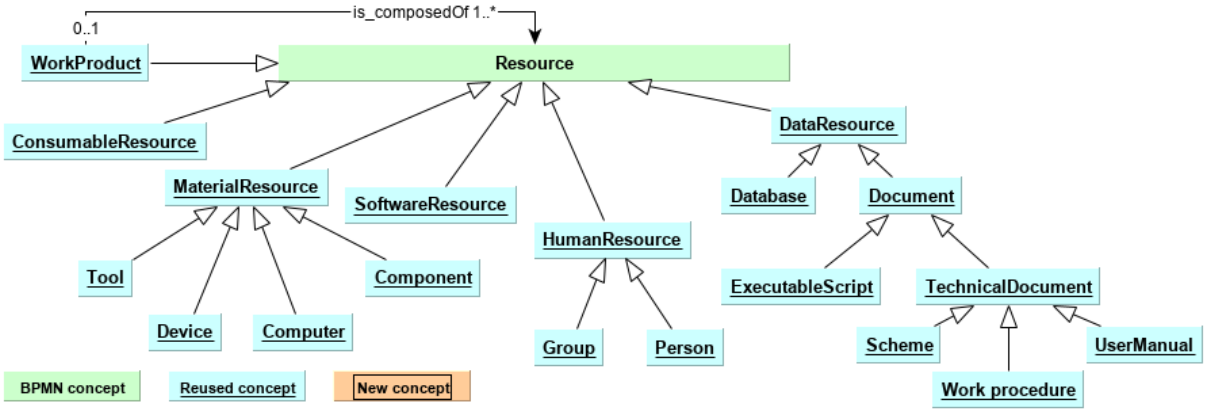


FIGURE *5 – Diagramme de classes correspondant à la taxonomie des ressources*

IC 2019

4.3 Site de production

Nous avons réutilisé la taxonomie proposée par (Chungoora *et al.*, 2013; Fraga *et al.*, 2018) pour décrire le site où une tâche doit être exécutée. Le concept racine *ManufacturingFacility* est spécialisé en cinq sous-classes liées par une relation de méronymie (voir Figure 6), du lieu le plus précis (la station de travail) au lieu le plus global (l'entreprise). Ainsi les sites de production peuvent être associés aux tâches, aux activités, etc.

Le niveau de description le plus fin demande à spécifier le site de production pour chaque tâche (un site par tâche élémentaire) grâce à la relation "takesPlaceAt". Or le modèle permet aussi de spécifier un site pour une activité ou un processus, pour les cas où toutes les tâches de l'activité ou du processus doivent être exécutées sur le même site.

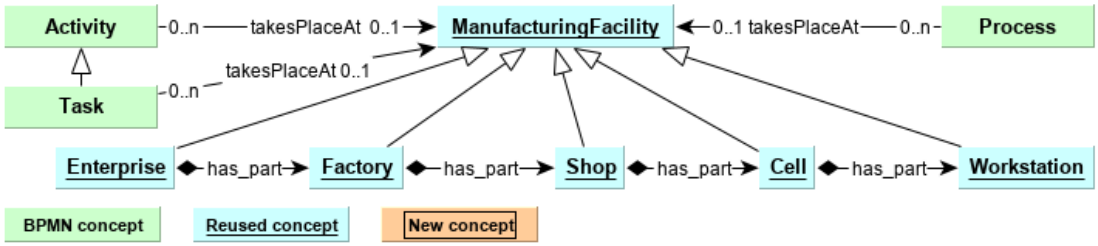


FIGURE 6 – *Diagramme UML du concept Site de production*

4.4 Produit manufacturé

Dans les normes ISO 10303-1 et 15531-32, le produit est défini par : "*Thing or substance produced by a natural or artificial process.*". La norme ISO 10303-239, quant à elle, définit la ressource comme suit : "*Resource is the result of a process.*" Ces définitions nous permettent de considérer le produit (*WorkProduct*) résultant de l'exécution d'un processus comme un groupe de ressources, à leur tour utilisables pour en produire d'autres. Cela est représenté à l'aide de la relation *is_composedOf* entre *WorkProduct* et *Resource* (Figure 5).

4.5 Agent

Le diagramme de classe de la Figure 7 est inspiré des travaux de (Ruíz *et al.*, 2004). Un agent peut être une ressource humaine ou logicielle. Pour une activité donnée, il est possible d'affecter un agent particulier (affectation directe), ou de spécifier son rôle (affectation indirecte).

Par ailleurs, les deux relations "subordinated" et "superior" liées au concept *Job* représentent le modèle organisationnel de l'entreprise, qui n'est pas décrit dans BPMN. La représentation de ce modèle est intéressante pour identifier les responsabilités hiérarchiques, par exemple lors d'un Workflow de validation ou de l'envoi de notifications en cas d'anomalie durant le déroulement d'un processus.

Nous avons différencié le concept *Job* du concept *Role* pour offrir plus de flexibilité. En effet, deux personnes ayant le même poste (Job) n'ont pas systématiquement les mêmes autorisations pour exécuter des activités.

5 Formalisation et implémentation en OWL

La formalisation de BBO s'est faite en deux temps : nous avons tout d'abord appliqué un ensemble de règles pour traduire les diagrammes de classes UML créés à la phase précédente en des concepts et relations OWL ; puis nous avons transcrit un ensemble de spécifications énoncées en langage naturel en OWL.

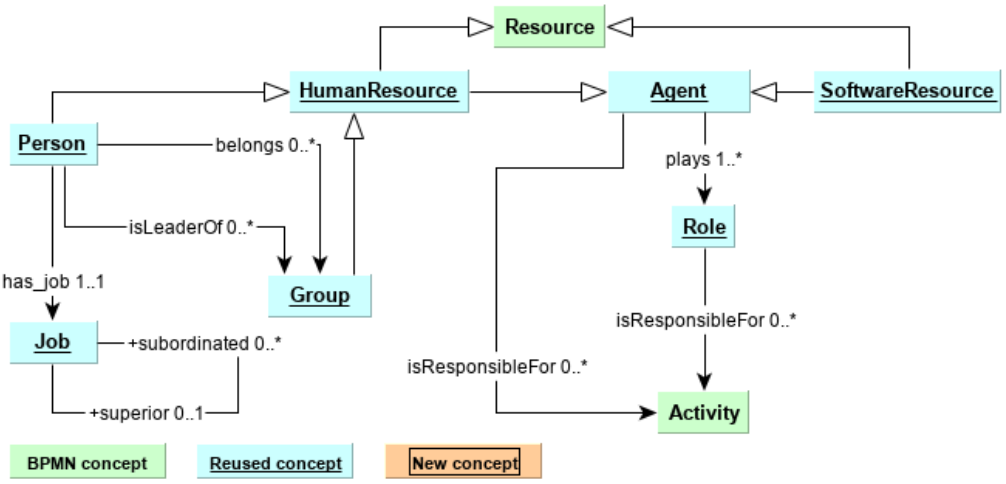


FIGURE 7 – *Diagramme UML du concept Agent*

5.1 Traduction des diagrammes UML en OWL

Nous avons appliqué les règles suivantes :

1. Chaque classe UML donne lieu à la création d'une classe OWL
2. Chaque relation entre classes UML est représentée par une propriété OWL de type ObjectProperty dont le domaine et le co-domaine sont les classes OWL correspondant à ces classes UML.
3. Pour chaque attribut de classe UML :
   — Si le type de cet attribut est une autre classe UML, une propriété OWL de type ObjectProperty est créée, avec pour domaine la classe à laquelle appartient l'attribut et pour co-domaine la classe associée au type de l'attribut.
   — Sinon, une propriété OWL de type DatatypeProperty est créée, liant la classe OWL créée à partir de la classe UML à laquelle est associé l'attribut au type de donnée XML spécifié dans le diagramme UML.
4. Les cardinalités sont traduites en restrictions de cardinalité sur les propriétés. Soient UClass1 et UClass2 deux classes UML, OClass1 et OClass2 les deux classes OWL correspondantes, UProperty une relation UML ou un attribut, et OProperty la propriété OWL correspondante. Soit n un entier et x un entier supérieur à 0.
   — UClass1 UProperty [x..x] UClass2 => OClass1 (subClassOf OProperty **exactly** x) OClass2
   — UClass1 UProperty [x..n] UClass2 => OClass1 (subClassOf OProperty **min** x) OClass2
   — UClass1 UProperty [0..x] UClass2 => OClass1 (subClassOf OProperty **max** x) OClass2

5.2 Traduction en OWL de spécifications exprimées en LN

L'équipe OMG, qui a développé le La définition du modèle BPMN par l'équipe OMG comporte, en plus du diagramme UML, un ensemble de spécifications rédigées en langage naturel dont une partie n'est pas prise en charge par le diagramme UML. Or, pour des questions de cohérence et de validation de la représentation des processus, il est important de formaliser le sous-ensemble de ces spécifications qui concernent les concepts de notre ontologie. Nous donnons quelques exemples dans la Table 4 (les formalisations issues des spécifications en langage naturel sont notées en caractères gras).

La formalisation de ces spécifications consiste à exprimer des restrictions sur des propriétés existantes (voir les 2 premiers exemples de la Table 4), ou encore à spécialiser des concepts existants (voir les 2 derniers exemples de la Table 4 où les concepts *Converging-Gateway* et *ConditionalSequenceFlow* sont donc de nouveaux sous-concepts de *Gateway* et

Spécification	**Formalisation du texte**
A Start Event **MUST** be a source for a Sequence Flow (p.245)	StartEvent *subClassOf* (CatchEvent *and* ... ***and*** **has_outgoing *some* SequenceFlow**)
The list of BPMN elements that **MUST NOT** be used in an Ad-Hoc Sub-Process : Start Event, End Event (p.182)	**AdHocSubProcess *SubClassOf not* (has_flowElements *some* (StartEvent or EndEvent))**
A Gateway with a gatewayDirection of converging MUST have multiple incoming Sequence Flows, but **MUST NOT** have multiple outgoing Sequence Flows. (p. 290)	**ConvergingGateway *equivalentTo*** **(Gateway** **and (has_incoming *min* 2 SequenceFlow)** ***and*** **(has_outgoing *exactly* 1 SequenceFlow))**
A Timer Event is an Event that has exactly one TimerEventDefinition. (p. 274)	**TimerEvent *equivalentTo*** **(Event** ***and*** **(has_eventDefinition *exactly* 1 TimerEventDefinition))**
An Intermediate Event **MUST** be a source for a Sequence Flow. (p. 259)	IntermediateEvent *equivalentTo* (IntermediateCatchEvent *or* IntermediateThrowEvent) **IntermediateEvent *subClassOf* (has_outgoing *some* SequenceFlow)**

TABLE *4 – Formalisation de spécifications de BPMN*

de *SequenceFlow* respectivement). Cette formalisation permet une classification automatique des instances et une vérification du respect des spécifications. BBO a été implémentée en OWL à l'aide de l'outil d'édition d'ontologie Protégé.

5.3 Documentation

Enfin, nous avons documenté BBO en ajoutant la description de chaque concept et de chaque relation à l'aide de la propriété *rdfs :comment*. Pour les concepts et relations issus des modèles réutilisés, les descriptions originales ont été respectées.

6 Évaluation et Discussion

Nous avons vu en Section 4 que le modèle sans doute le plus abouti pour représenter un processus métier était le modèle BPMN, mais nous avons aussi souligné ses limites. Ne disposant donc pas d'ontologie de référence pour évaluer BBO, nous proposons d'estimer sa qualité selon des critères quantitatifs et qualitatifs.

6.1 Évaluation quantitative

Afin d'évaluer le modèle conceptuel de BBO, nous avons calculé deux métriques proposées par (Tartir & Arpinar, 2007). Soient NH le nombre de relations d'hyperonymie (is-a), NR le nombre de relations autres que l'hyperonymie, et NC le nombre de concepts. Les indicateurs sont ainsi calculés :

— RD = NR/(NR + NH) RD est compris entre 0 et 1, et il indique le taux de diversité des relations. Plus RD est proche de 1, plus les relations sont diverses au sein du modèle.
— SD = NH / NC indique la profondeur de l'ontologie et correspond au nombre moyen de relations d'hyperonymie entre chaque concept et la racine. Un SD faible traduit le fait que l'ontologie est "profonde" et détaille les connaissances d'un domaine spécifique.

La Table 5 donne le nombre actuel de concepts et de relations dans le modèle conceptuel de BBO, ainsi que les valeurs des deux métriques RD et SD.

#Concepts	#Relations autres que is-a	#Relations is-a	Métriques
158	145	135	RD = 145/(145+135)= 0.52 SD = 135/158= 0.85

TABLE *5 – Métriques du schéma BBO*

La valeur de RD est 0.52, ce qui montre que BBO n'est pas qu'une simple taxonomie, mais une ontologie riche en relations. De plus, BBO a un SD faible inférieur à 1, ce qui témoigne de sa profondeur et d'une bonne couverture du domaine.

6.2 Évaluation qualitative

L'évaluation qualitative a été menée de façon empirique pour vérifier que BBO est bien (i) consistante ; (ii) expressive, i.e. capable de représenter différents modèles de processus métiers ; et (iii) capable de répondre aux questions portant sur les compétences techniques.

6.2.1 Consistance

Nous avons vérifié la consistance de BBO en utilisant trois raisonneurs HermiT, Fact et Pellet au sein de Protégé, et cela avant et après son peuplement.

6.2.2 Expressivité

Nous avons peuplé BBO avec deux processus métier, choisis aléatoirement, et provenant des deux partenaires industriels, TAS et Continental. Pour des raisons de confidentialité, nous n'allons présenter que des extraits anonymes des processus originaux.

Exemple 1. : Ce premier exemple concerne l'extrait du processus TAS présenté en Table 1. Pour plus de lisibilité, nous représentons l'exemple avec des éléments graphiques du langage graphique de BPMN, et nous affectons un identifiant (en rouge) à chaque élément (Figure 8).

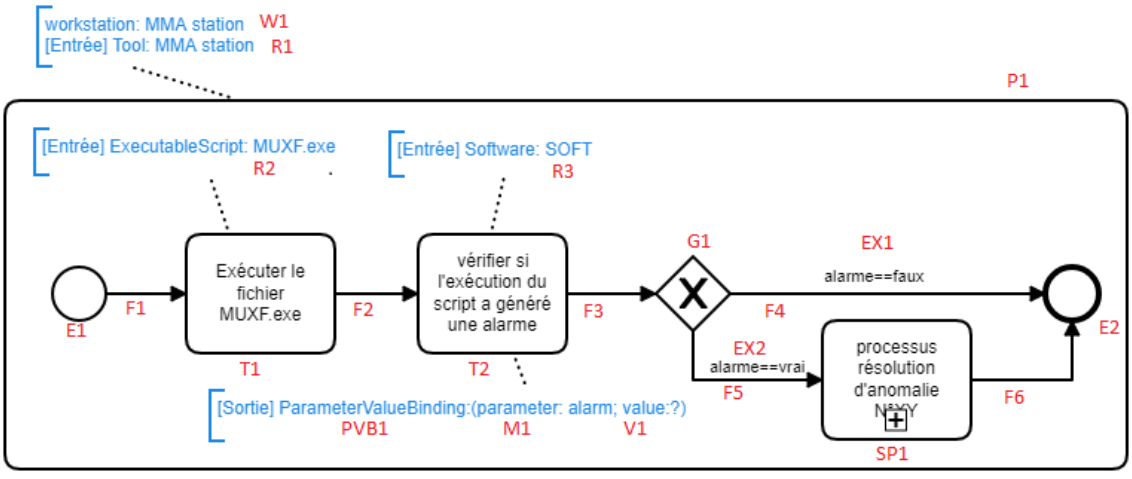


FIGURE *8 – Représentation graphique de l'extrait présenté dans la Table 1*

Le premier cercle est un *StartEvent*, point de démarrage du processus. Les deux premiers rectangles représentent les deux premières tâches. Au terme de la deuxième tâche, deux cas

sont possibles selon la valeur de la variable *alarme*, d'où l'utilisation de la passerelle (Gateway) représentée par un losange. Si la condition (alarme==vrai) est vraie, le sous processus (rectangle avec le symbole "+") doit être exécuté avant de terminer le processus. Le dernier cercle avec une bordure épaisse est un *EndEvent*, qui marque de la fin du processus.

Toutes les instances d'une classe OWL sont de type `owl:NamedIndividual`. La Table 6 donne les triplets qui lient les instances aux classes de BBO, en utilisant les identifiants de la Figure 8. Les assertions sont écrites en Turtle avec un espace de nommage de base correspondant à l'URI de BBO (pour simplifier, nous avons omis le préfixe BBO e.g., `:Process` au lieu de `BBO:Process`). La tâche T2 a une spécification de sortie qui est la valeur V1 du paramètre M1 (paramètre alarme), représenté en créant une instance PVB1 de *ParameterValueBinding* qui relie le paramètre à sa valeur qui va être renseignée durant l'exécution.

La table 7 montre les assertions correspondant aux instances des relations de type ObjectProperty. D'autres assertions correspondant aux relations de DataProperty seraient nécessaires pour une représentation complète, mais nous les omettons pour des raisons de lisibilité.

:P1 a :Process.	:R3 a :Software .
:E1 a :StartEvent .	:M1 a :Parameter .
:Fi a :SequenceFlow . #(i=1,..,6)	:V1 a :ParameterValue .
:Ti a :Task #(i=1,2)	:PVB1 a :ParameterValueBinding .
:G1 a :Gateway .	:W1 a :Workstation .
:IOi a :InputOutputSpecification #(i=1,...,3)	:EXi a :ConditionalExpression #(i=1,2)
:R1 a :Tool .	:SP1 a :SubProcess .
:R2 a :ExecutableScript .	:E2 a :EndEvent .

TABLE *6 – Triplets définissant les instances des classes OWL*

:P1 takesPlaceAt :W1 .	:EX1 :has_valueToBeEvaluated :V1.
:P1 :has_flowElements :E1.	:EX2 :has_valueToBeEvaluated :V1.
:P1 :has_flowElements :F1, :F2, :F3, :F4, :F5, :F6.	:F1 :has_sourceRef :E1.
:P1 :has_flowElements :Ti i=1,2 .	:F1 :has_targetRef :T1.
:P1 :has_flowElements :G1 .	:F2 :has_sourceRef :T1.
:P1 :has_ioSpecification :IO3 .	:F2 :has_targetRef :T2.
:IO3 :has_resourceInput :R1 .	:F3 :has_sourceRef :T2.
:P1 :has_flowElements :SP1 .	:F3 :has_targetRef :G1.
:P1 :has_flowElements :E2 .	:F4 :has_sourceRef :G1.
:IO1 :has_resourceInput :R2 .	:F4 :has_targetRef :E2.
:T1 :has_ioSpecification :IO1 .	:F4 :has_conditionExpression :EX1.
:IO2 :has_resourceInput :R3 .	:F5 :has_sourceRef :G1.
:IO2 :has_parameterValueBindingOutput :PVB1 .	:F5 :has_targetRef :SP1.
:PVB1 :has_parameter :M1 .	:F5 :has_conditionExpression :EX2.
:PVB1 :has_parameterValue :V1 .	:F6 :has_sourceRef :SP1.
:T2 :has_ioSpecification :IO2 .	:F6 :has_targetRef :E2.

TABLE *7 – Axiomes correspondant aux propriétés OWL*

Exemple 2. : La Table 8 présente un autre exemple fourni par Continental. L'instruction 45 précise que les tâches 41 à 44 doivent être répétées 9 fois. BBO permet la représentation de cette spécification. En effet, il suffit de créer une instance SLC de *StandardLoopCharacteristics* qui est une sous-classe de *LoopCharacteristics* (Figure 2). *StandardLoopCharacteristics* a un attribut, appelé *loopMaximum*, qui permet de spécifier le nombre max d'itérations. Pour cet exemple, on affecte la valeur 9 à l'attribut *loopMaximum* de SLC. Ensuite, il faut créer une instance de *SubProcess* (SP) qui inclut les tâches 41 à 44, et la lier à SLC par la propriété *has_loopCharacteristics* (Table 9).

```
43. ...................
44. ...................
45. Répéter les opérations 41 à 44, 9 fois en suivant
les instructions sur l'écran du poste d'emballage
```

TABLE *8 – Extrait anonymisé d'un processus métier Continental*

:SP a :SubProcess.	:SLC :loopMaximum 9.
:SP :has_flowElements :T41, :T42, :T43, :T44.	:SP :has_loopCharacteristics :SLC.
:SLC a :StandardLoopCharacteristics.	

TABLE *9 – Les assertions de l'exemple 2*

Exemple 3. Dans cet exemple, nous mettons l'accent sur la quatrième instruction de l'exemple présenté Table 10. L'opérateur qui l'exécute est censé saisir la valeur de la télémesure TM_MM et vérifier si sa valeur est comprise entre -0.8V et +0.8V. La représentation de cette instruction fait appel aux concepts et relations que nous avons rajoutés pour représenter les paramètres d'entrée/sortie d'une activité (Figure 4). En effet, la télémesure qu'envoie le satellite est un paramètre qui prend différentes valeurs selon la télécommande envoyée.

Comme indiqué dans la Table 11, la télémesure TM_MM est représentée comme une instance de la classe *QuantitativeParameter*. Pour spécifier l'unité de mesure, nous rajoutons une instance de la classe *UnitOfMeasure* qui référence via la proprité *has_unit* une instance de la classe *UO_0000218* (la classe des Volts de l'ontologie des unités de mesure UO, importée dans BBO). Puis, nous créons EV, une instance de *ParameterExpectedValue* dont les valeurs min et max sont spécifiées par les attributs *minValue* et *maxValue*, à savoir -0.8 et +0.8. Nous créons aussi une instance de *ParameterValue* (PV) pour spécifier que T4 va produire une valeur durant l'exécution du processus. PVB est une instance de *ParameterValueBinding* qui lie le paramètre à ses valeurs, à l'aide des propriété *has_parameter*, *has_parameterValue* et *has_parameterExpectedValue*. Enfin, nous rajoutons une instance de *InputOutputSpecification* qui fera la liaison entre T4 et PVB grâce aux priopriétés *has_ioSpecification* et *has_parameterVaueBindingOutput*.

```
1. ...................
2. Envoyez la télécommande CGHFR.
3. Attendre 45 secondes.
4. Vérifier et noter la valeur de la télémesure
suivante TM_MM (-0.8V,+0.8V)
```

TABLE *10 – Extrait anonymisé d'un processus métier TAS*

Ainsi, nous avons pu vérifier que BBO permet de représenter les connaissances contenues dans les deux types de processus métier traités.

6.2.3 Questions basées sur les compétences techniques

Nous avons pu exprimer toutes les "competency questions" ayant servi à spécifier le modèle (Section 3.2) en SPARQL, et avons exécuté ces requêtes sur BBO peuplée, dans l'environnement Protégé (Table 12).

Pour certaines questions, il n'était pas possible de vérifier les résultats des requêtes : (i) soit parce qu'il n'y a pas d'instance dans la base de connaissances qui réponde à la requête ; c'est le cas de la question 7 car les fiches décrivant les processus métiers ne mentionnent pas les agents ou les rôles responsables de l'exécution des activités ; (ii) soit parce que la

:T4 a :Task.	:EV :maxValue +0.8.
:TM_MM a :QuantitativeParameter.	:PV a :ParameterValue.
:UM a :UnitOfMeasure.	:PVB a :ParameterValueBinding.
:Volt a UO :UO_0000218.	:PVB :has_parameter :TM_MM.
:UM has_unit :Volt.	:PVB :has_parameterValue :PV.
:TM_MM has_unitOfMeasure :UM.	:PVB :has_parameterExpectedValue :EV.
:IO a :InputOutputSpecification.	:IO :has_parameterValueBindingOutput :PVB.
:EV a :ParameterExpectedValue.	:T4 :has_ioSpecification :IO.
:EV :minValue -0.8.	

TABLE *11 – Les assertions de l'exemple 3*

réponse relève de l'aspect dynamique de ces processus métier ; la question 5 par exemple cherche l'activité qui suit une activité donnée. Cela dépend parfois de l'évaluation de certains paramètres durant l'exécution, tel que le paramètre "alarme" dans la Figure 8.

6.2.4 Synthèse

Cette évaluation a montré que BBO est riche en relations autres que "is-a", et que son niveau de profondeur offre une bonne couverture du domaine. BBO permet ainsi de représenter les processus métiers avec une granularité fine, comme nous l'avons observé à travers le peuplement avec les deux processus. Enfin, on était capable de formaliser toutes les competency questions à l'aide de BBO. Sur la base de cette évaluation empirique, nous estimons que BBO présente les qualités requises pour répondre aux besoins de l'agent virtuel en termes de description des processus métier. BBO est accessible à partir de l'URL `https://www.irit.fr/recherches/MELODI/ontologies/BBO`.

7 Conclusion et Perspectives

La modélisation des processus métier est un domaine actif de recherche qui attire plus d'attention avec l'émergence de l'industrie 4.0. Dans cet article, nous avons proposé une nouvelle ontologie, appelée BBO, pour représenter les processus métier. Pour construire cette ontologie, nous avons réutilisé des fragments d'ontologies et de modèles existants, notamment BPMN 2.0. L'évaluation que nous avons menée a montré que BBO (i) n'est pas qu'une simple taxonomie, (ii) n'a pas une structure plate, et peut donc décrire avec finesse les connaissances du domaine, (iii) est une ontologie consistante, (iv) est capable de représenter formellement les "competency questions" issues des besoins exprimés par les experts et de la littérature. BBO a été implémentée en OWL et mise à disposition de la communauté scientifique.

Sur la base de cette évaluation, BBO permet de représenter les informations nécessaires à l'exécution des processus métiers. Par contre, BBO n'assure pas encore la représentation de l'historique des exécutions. Ce point est la prochaine étape de notre travail. Par ailleurs, peupler l'ontologie manuellement est une besogne longue et fastidieuse, et sujette aux erreurs. Nous projetons de la réaliser automatiquement ou semi-automatiquement, par un traitement automatique du texte présent dans les documents techniques décrivant les processus métiers.

8 Remerciements

Le projet AVI-REX est financé par la Région Occitanie, le FEDER-FSE Midi-Pyrénées et le programme Garonne 2014-2020 sous le label 2017-AVIREX-IRIT-READYNOV INDUSTRIE DU FUTUR. Les auteurs remercient leurs partenaires industriels, SIMSOFT INDUSTRY, Thales Alenia Space et Continental.

N°	Question en langage naturel	Requête en SPARQL
1	`Quelles ressources sont requises pour une activité ?` -Plusieurs versions de cette question peuvent être posées selon la nature de la ressource (ex : les outils nécessaires, les composants nécessaires, etc.)	Select ?A ?R where { ?A a BBO :Activity. ?R a BBO :Resource. ?A BBO :has_ioSpecification ?io. ?io BBO :has_resourceInputs ?R}
2	`Quelles sont les activités qui composent un processus ?` - Cette question peut aussi être posée pour un sous-processus.	Select ?P ?A where { ?P a BBO :Process. ?A a BBO :Activity. ?P BBO :has_flowElements ?A}
3	`Quelles activités doivent être achevées avant de commencer une activité donnée ?` - Cette question traite la dépendance entre les activités. Une activité A dépend d'une activité B, si B produit une sortie (ressource ou valeur de paramètre) qui est l'entrée de A. La requête suivante renvoie tous les couples d'activités (A,B) tels que A dépend de B.	Select Distinct ?A ?B where {{ ?B BBO :has_ioSpecification ?IOb. ?IOb BBO :has_resourceOutputs ?R. ?A BBO :has_ioSpecification ?IOa. ?IOa BBO :has_resourceInputs ?R.} UNION { ?B BBO :has_ioSpecification ?IOb. ?IOb BBO :has_parameterValueBindingOutputs ?PVB. ?A BBO :has_ioSpecification ?IOa. ?IOa BBO :has_parameterValueBindingInputs ?PVB.}}
4	`Quel est le site d'exécution d'une activité ?`	Select ?A ?S where { ?A a BBO :Activity. ?S a BBO :ManufacturingFacility. ?A BBO :takesPlaceAt ?S.}
5	`Quelle est l'activité qui suit une activité ?` -La réponse à cette requête n'est pas toujours simple. Par exemple, dans la Figure 8, l'activité qui suit T2 ne peut être connue qu'au moment de l'exécution car elle dépend de la valeur de « alarme » que l'opérateur doit saisir. Dans le cas de T2, la requête va donner un ensemble vide, mais il faut chercher la nature de l'élément suivant Event, Gateway ou Activity pour continuer l'exécution du processus.	Select ?A ?nextA where { ?A a BBO :Activity. ?nextA a BBO :Activity. ?A BBO :has_outgoing ?SF. ?SF BBO :has_targetRef ?nextA.}
6	`Quel est le type de chaque ressource ?` - La taxonomie des ressources de BBO permet de donner un type précis à chaque ressource. Cette taxonomie peut être spécialisée pour chaque entreprise afin de prendre en compte ses ressources. Ainsi, on a ajouté `Switch` comme sous-classe de `Component` pour TAS et `LecteurCodeBar` comme sous-classe de `Device` pour Continental.	Select Distinct ?R ?typeR where { ?R a BBO :Resource. ?R a ?typeR.}
7	`Qui peut exécuter chacune des activités ?` -l'affectation des agents aux différentes activités peut se faire directement ou par rôle, d'où l'union des deux requêtes.	Select ?A ?agent where { { ?agent a BBO :Agent. ?A a BBO :Activity. ?agent BBO :isResponsibleFor ?A.} UNION { ?agent a BBO :Agent. ?A a BBO :Activity. ?role a BBO :Role. ?agent BBO :has_role ?role. ?role BBO :isResponsibleFor ?A.}}
8	`Quels sont les composants de chaque produit ?` -il est possible de répondre à cette question grace à la relation isComposedof entre WorkProduct et Resource	Select ?w ?r where { ?w a BBO :WorkProduct. ?r a BBO :Resource. ?w BBO :isComposedOf ?r}
9	`Y-a-t-il un manuel d'utilisateur pour un outil donné ?` -la requête renvoie la liste de tous les outils qui ont un manuel d'utilisateur, il suffit de filtrer par le code ou le nom de l'outil spécifié pour avoir la réponse.	Select ?T, ?UM where { ?T a BBO :Tool. ?UM a BBO :UserManual. ?T BBO :has_technicalDocument ?UM}
10	`What is the unit measure of a given parameter?` - l'unité de mesure d'un paramètre se compose d'une partie principale (unité) et eventuellement un prefix (mili, centi, etc.), d'où le mot clé optionnal pour le prefix.	SELECT ?P ?unit ?prefix WHERE { ?P a BBO :QuantitativeParameter. ?P BBO :has_unitOfMeasure ?unit. optional { ?P BBO :has_prefix ?prefix} }

TABLE *12 – Exemples de questions basées sur les compétences techniques en SPARQL.*

Références

ABDALLA A., HU Y., CARRAL D., LI N. & JANOWICZ K. (2014). An Ontology Design Pattern for Activity Reasoning. In *5th Workshop on Ontology and Semantic Web Patterns (WOP2014)*, p. 78–81, Riva del Garda, Italy.

AWAD A., GROSSKOPF A., MEYER A. & WESKE M. (2009). *Enabling resource assignment constraints in BPMN*. Rapport interne, Hasso Plattner Institute.

BERNERS-LEE T., HENDLER J. & LASSILA O. (2001). The Semantic Web. *Scientific American*, **284**(5), 28–37.

BOCCIARELLI P., D'AMBROGIO A., GIGLIO A. & PAGLIA E. (2016). A BPMN extension to enable the explicit modeling of task resources. In *CEUR Workshop Proceedings*, volume 1728, p. 40–47.

CABRAL L., NORTON B. & DOMINGUE J. (2009). The business process modelling ontology. In *4th international workshop on semantic business process management*, p. 9–16.

IC 2019

CHUNGOORA N., YOUNG R. I. M., GUNENDRAN G., PALMER C., USMAN Z., ANJUM N. A., CUTTING-DECELLE A.-F., HARDING J. A. & CASE K. (2013). A model-driven ontology approach for manufacturing system interoperability and knowledge sharing. *Computers in Industry*, **64**(4), 392–401.

FALBO R. D. A. & BERTOLLO G. (2009). A software process ontology as a common vocabulary about software processes. *International Journal of Business Process Integration and Management*, **4**(4), 239–250.

FERNÁNDEZ M., GÓMEZ-PÉREZ A. & JURISTO N. (1997). METHONTOLOGY : From Ontological Art Towards Ontological Engineering. In *Symposium on Ontological Engineering of American Asociation for Artificial Intelligence (AAAI)*, p. 33–40.

FRAGA A., VEGETTI M. & LEONE H. (2018). Semantic interoperability among industrial product data standards using an ontology network. In *20th International Conference on Enterprise Information Systems (ICEIS)*, volume 2, p. 328–335, Madeira, Portugal.

GRÜNINGER M. & FOX M. S. (1995). The Role of Competency Questions in Enterprise Engineering. In *Benchmarking—Theory and practice*, p. 22–31. Springer, Boston, MA.

KARRAY M. H., CHEBEL-MORELLO B. & ZERHOUNI N. (2012). A formal ontology for industrial maintenance. *Applied ontology*, **7**(3), 269–310.

MARCINKOWSKI B. & KUCIAPSKI M. (2012). A business process modeling notation extension for risk handling. In *11th International Conference on Computer Information Systems and Industrial Management (CISIM)*, volume 7564 LNCS, p. 374–381, Venice, Italy.

NATSCHLÄGER C. (2011). Towards a BPMN 2.0 ontology. In *3rd International Workshop on Business Process Modeling Notation*, p. 1–15, Lucerne, Switzerland.

OMG (2011). *Business Process Modeling Notation, v2.0 - Specification*. Rapport interne, Object Management Group.

PEDRINACI C., DOMINGUE J. & DE MEDEIROS A. K. A. (2008). A core ontology for business process analysis. In *5th European Semantic Web Conference, ESWC, Tenerife, Canary Islands, Spain*, p. 49–64.

ROSPOCHER M., GHIDINI C. & SERAFINI L. (2014). An ontology for the Business Process Modelling Notation. In *8th International Conference on Formal Ontology in Information Systems (FOIS)*, p. 133–146, Rio de Janeiro, Brazil.

ROY S., DAYAN G. S. & DEVARAJA HOLLA V. (2018). Modeling industrial business processes for querying and retrieving using OWL+SWRL. In *On the Move to Meaningful Internet Systems (OTM)*, p. 516–536, Valletta, Malta.

RUÍZ F., VIZCAINO A., PIATTINI M. & GARCÍA F. (2004). An Ontology For The Management Of Software Maintenance Projects. *International Journal of Software Engineering and Knowledge Engineering*, **14**(3), 1–27.

STROPPI L. J. R., CHIOTTI O. & VILLARREAL P. D. (2011). A BPMN 2 . 0 Extension to Define the Resource Perspective of Business Process Models. In *14th Iberoamerican Conference on Software Engineering (CIbSE)*, p. 25–38, Rio de Janeiro, Brasil.

TARTIR S. & ARPINAR I. B. (2007). Ontology evaluation and ranking using OntoQA. In *1st International Conference on Semantic Computing (ICSC)*, p. 185–192, California, USA.

USCHOLD M., KING M., MORALEE S. & ZORGIOS Y. (1998). The enterprise ontology. *The knowledge engineering review*, **13**(1), 31–89.

VOGEL-HEUSER B. & HESS D. (2016). Guest Editorial Industry 4.0–Prerequisites and Visions. *IEEE Transactions on Automation Science and Engineering*, **13**(2), 411–413.

Trois conceptions du processus : les raisons d'un choix

Gilles Kassel

Laboratoire MIS, Université de Picardie Jules Verne
33 rue Saint-Leu, 80039 Amiens Cedex 1
Gilles.kassel@u-picardie.fr

Résumé : L'ontologie des entités qualifiées d'« occurrentes » – *processus*, *états* et *événements* – est un domaine actif de recherche auquel nous tâchons de contribuer. Nous avons ainsi proposé récemment un cadre ontologique combinant des processus *endurants* (plutôt qu'*occurrents*) et des événements *abstraits* (plutôt que *concrets*). L'ontologie des processus, plus particulièrement, continue de connaître des développements, comme en témoigne l'ouvrage édité en 2018 par Rowland Stout *Process, Action, and Experience*. Toutefois, un constat notoire est qu'aucun consensus ne se dégage sur la nature des processus. Au contraire, trois figures du processus sont actuellement promues : (1) le processus comme continuant dynamique, proche de l'objet dans sa façon d'endurer dans le temps ; (2) le processus comme occurrent étendu temporellement et constituant la « matière » d'événements ; enfin, (3) le processus comme forme temporelle abstraite réalisée par des occurrences concrètes. Dans cet article, nous approfondissons les bases conceptuelles et ontologiques de ces figures. En résultat, nous argumentons en faveur de la figure (1) et en défaveur de la figure (2). Par ailleurs, nous montrons que la figure (3) est compatible avec la figure (1), mais précisons qu'elle caractérise le *mouvement* plutôt que le *processus*.

Mots-clés : Ontologie formelle, processus physique, événement, changement, changement de Cambridge, changement réel, mouvement

1 Introduction

Récemment, nous avons entamé un travail consistant à élaborer une nouvelle ontologie des entités dites « occurrentes », parmi lesquelles on distingue couramment les *processus*, *états* et *événements* (Kassel, 2017, 2018). Le cadre ontologique que nous proposons repose sur trois thèses principales : (i) le monde physique est peuplé de particuliers – *objets* et *processus* – qui, en endurant, assurent sa stabilité tout autant que sa dynamique ; (ii) ces particuliers portent temporairement des *propriétés*, ces exemplifications de propriétés correspondant à des *faits* existant dans le monde physique ; par ailleurs, (iii) des sujets cognitifs, plongés dans le monde physique, se représentent au moyen d'*événements* l'histoire (passée, présente et future) du monde pour interagir avec lui. Avec ce cadre, nous ne cherchons pas à édifier une ontologie de l'*objet* en général ou du *processus* en général, mais nous contentons de caractériser les *objets physiques* et les *processus physiques*. L'inventaire que nous dressons des entités « occurrentes » diffère de l'inventaire courant en ontologie formelle comme en ontologie appliquée. Nous distinguons ainsi, d'une part, les *processus* (mais ceux-ci *endurent* plutôt qu'ils *occurrent*), d'autre part, les *faits dynamiques* (par exemple les faits de *perpétuation* de processus) et, enfin, les *événements* existant pour des sujets et qui occurrent lorsqu'ils sont réalisés par des processus (Kassel, 2019).

La figure ontologique du processus physique que nous retenons relève de théories récentes et encore minoritaires en ontologie formelle. L'ontologie des processus continue d'être un domaine actif de travaux, en témoigne l'ouvrage collectif édité par Rowland Stout (2018). Par contre, le constat notoire troublant est que les développements proposés ne s'accordent pas

autour d'une même figure du processus. Au contraire, ils mettent en scène principalement trois conceptions a priori opposées : (1) le processus comme continuant dynamique, proche de l'objet dans sa façon d'endurer dans le temps (Stout, 1997, 2003, 2016 ; Galton, 2006, 2008) ; (2) le processus comme occurrent étendu temporellement, constituant la « matière » des événements (Mourelatos, 1978 ; Galton et Mizoguchi, 2009 ; Crowther, 2011, 2018) ; enfin, (3) le processus comme pattern abstrait temporel, réalisé par des occurrents concrets (Galton, 2012, 2018). Dans nos travaux, nous promouvons la figure (1) mais, notamment du fait de l'ontologie du temps que nous adoptons, nous retenons une classe de processus plus restreinte que celle considérée par Galton et Stout.

Dans cet article, notre objectif est d'approfondir les bases conceptuelles et ontologiques de ces trois figures et de dégager les raisons du choix que nous avons effectué, renforçant par là-même notre cadre ontologique. Pour mener cette analyse, le fil conducteur que nous suivons est l'analyse du *changement* qualitatif, à savoir le fait pour une substance de porter des propriétés contradictoires à différents temps, et plus particulièrement l'analyse du *mouvement*, en tant que changement de localisation spatiale. Nous allons tout d'abord en Section 2 rappeler la contribution d'Aristote à l'analyse du changement, qui est l'une des premières dont nous disposions et qui contient en germes les théories les plus récentes. En visant à justifier la réalité du changement, un apport considérable d'Aristote a été de dégager le *processus* en proposant une analyse *dispositionnelle*, toujours d'actualité. En Section 3, nous effectuons un saut de plus de 2000 ans dans le temps avec la conception du changement promue au début du 20^{e} siècle par des philosophes de Cambridge (McTaggart, Russell) et nommée pour cette raison *changement de Cambridge*. Nous en rappelons les critiques et présentons l'analyse dispositionnelle du processus proposée par Cleland (1990), visant à définir le changement « réel ». En adoptant cette analyse, révisant celle d'Aristote, nous distinguons le 'processus du changement' du 'changement' lui-même. Ce préambule étant établi, en Section 4 nous évaluons les figures du processus (1), (2) et (3) précitées. En résultat, nous argumentons en faveur de la figure (1) et en défaveur de la figure (2). Par ailleurs, nous montrons que la figure (3) est compatible avec la figure (1), mais précisons qu'elle caractérise le *mouvement* (et, plus généralement, le *changement*) plutôt que le *processus du changement*.

2 Le traitement du changement chez Aristote

Aristote nous livre sa conception du changement dans son ouvrage *Physique*. Ses motivations sont alors doubles : 1) défendre la *réalité* du changement en réponse à Parmenide (et plus largement à l'école d'Élée) niant l'existence du changement ; et 2) élucider la *nature* du changement, là où Platon s'était contenté d'en décrire les causes et différentes formes[1].

La réalité du changement, Aristote entreprend de la défendre avec sa théorie de l'*être* (ou de la *substance*) comme composé de *matière* et d'une *forme*. La matière est ce qui permet à l'être de subsister dans le temps. Quant à la forme, celle-ci détermine la nature (ou *essence*) de l'être. Elle consiste en des attributs (ou *qualités*) dont certains peuvent varier en leur contraire. Ceci explique qu'une substance puisse se trouver dans des états différents pour une même qualité : un être peut être tantôt froid ou chaud, tantôt blanc ou noir et se trouver tantôt en un lieu ou un autre. Aristote distingue trois espèces de changements que sont l'accroissement ou la diminution en taille d'un corps, le changement de qualité (l'altération) et le changement de lieu (locomotion)[2].

Le changement (ou mouvement[3]) est par nature, et de façon générale, l'*acte du possible*, c'est-à-dire la réalisation d'un possible existant en tant que potentialité dans l'être (dans sa

[1] Ces motivations sont détaillées par Jules Barthélemy-Saint-Hilaire dans la longue préface de son (1862) *Physique d'Aristote*, première traduction française de l'ouvrage d'Aristote.

[2] Ce faisant, il écarte la génération et la destruction d'un être au prétexte que, dans de tels phénomènes, aucune substance ne subsiste. Par contre, il reconnaît qu'un changement puisse être *partiel* ou *absolu* suivant que la substance concernée soit affectée en partie ou en totalité.

[3] Chez Aristote, aussi bien que chez ses commentateurs, les deux termes sont utilisés de façon synonyme.

forme). Les êtres naturels sont ainsi mus par eux-mêmes. Par exemple, les corps pesants tombent comme une pierre, tandis que les corps légers s'élèvent comme le feu. Dans ces deux cas, selon Aristote, ces corps comportent en privation leur « lieu naturel » vers lequel ils se dirigent. Ces changements consistent en la réalisation de ce lieu naturel. Par ailleurs, les êtres (naturels ou non) peuvent changer du fait d'une cause qui leur est externe, comme dans le cas de phénomènes d'agence. Des exemples favoris d'Aristote sont le fait, pour un morceau de bronze dans l'atelier d'un sculpteur, de devenir une statue, et le fait, pour des pierres et du mortier, de devenir une maison. Pour rendre compte de ces changements, naturels ou non, Aristote propose la définition suivante (dans son livre III, et selon la traduction d'Ursula Coope (2009)) : *Change is the actuality of what is potentialy in some particular different state, qua such*.

Aristote propose ainsi une analyse dispositionnelle du changement avec les notions d'*actualité* et de *potentialité* (en des termes contemporains, on parle d'*exercice* d'une *disposition*). Ces notions comportent notoirement des parts d'ombre, en témoignent les diverses traductions, interprétations et amendement fournis par les commentateurs d'Aristote, y compris les plus contemporains. Pour éclairer ces notions, nous retenons principalement l'analyse de Coope (2009)[4], que nous modulons avec celle d'Andreas Anagnostopoulos (2010).

Commençons par la *potentialité*. Selon ces auteurs, cette notion est à entendre au sens d'un potentiel d'une substance à être dirigé vers un état différent, que cet état corresponde à une qualité (potentialité à être froid) ou une substance (potentialité à être une statue). Ceci suppose que la substance porte en elle cet état différent. Cette caractérisation entraîne que la potentialité disparaisse dès lors que la substance actualise le nouvel état.

L'*actualité*, pour sa part, est à entendre au sens où un potentiel d'une substance à être dans un état différent est *actuel*. L'actualité permet de distinguer deux situations dans lesquelles se trouve une substance, à savoir : (1) le fait de posséder – de façon dormante – un potentiel à être dans un état différent, et (2) le fait que s'exerce – de façon active – ce potentiel. L'actualité est ainsi une activité exercée par une substance.

Si l'on s'en tient à cette caractérisation du changement, la question qui se pose est de savoir si Aristote a effectivement réussi à défendre l'existence du changement et à en préciser la nature. Répondre à cette question suppose que l'on se donne des critères de succès. Il serait ici peu pertinent de passer la théorie d'Aristote au crible des critères contemporains de l'ontologie formelle. Au lieu de cela, tâchons d'évaluer l'apport d'Aristote. En continuant à nous appuyer sur l'analyse de nos deux commentateurs de référence, nous apportons de fait une réponse nuancée : sur un plan ontologique (celui de l'existence), Aristote pose des conditions nécessaires et suffisantes pour l'existence du changement, mais ne propose pas de critère d'identité satisfaisant ; sur un plan métaphysique (celui de la nature), la notion de 'potentiel à être dans un état final donné' est, pour le moins, imprécise.

Sur un plan ontologique, fondamentalement (selon Anagnostopoulos (2010)), l'apport d'Aristote est de proposer avec la notion de potentialité une explication causale du devenir : se limiter (comme c'était le cas des prédécesseurs d'Aristote) à considérer pour une substance deux états '*être*' ou '*ne pas être*' et à constater des différences ne permet pas de rendre compte du *devenir* des substances, par exemple d'expliquer qu'un gland devienne chêne ou qu'un caillou projeté en l'air par un promeneur retombe par terre ; selon Aristote, le gland et le caillou possèdent respectivement une *potentialité à être un chêne* et une *potentialité à être sur le sol*. Il reste à toutefois à préciser la nature de cette potentialité, à définir un critère d'identité, et c'est là que le bât blesse.

Une des raisons notamment (selon Coope (2009)) tient à l'analyse des phénomènes d'agence. Aristote, en effet, identifie le *changement d'une substance* à l'*action de changer cette substance*[5].

[4] Une version proche a notamment été donnée par Aryeh Kosman dans son (1969) *Aristotle's Definition of Motion*.

[5] Coope (2009, note 5) : *In fact, Aristotle says that the agent's action is one and the same change as the change undergone by the thing acted upon, though it is different 'in being'. For example, 'building a house' and 'becoming a house' are one and the same change (the change in the bricks and mortar that is directed towards being a house); building a house is that change considered from the perspective of the agent; becoming a house is that change considered from the perspective of the thing acted-upon. They must, he thinks, be one and the same change,*

Une telle identification conduit à questionner la nature du potentiel et, tout particulièrement, la confusion régnant entre le physique et le mental.

Une seconde raison tient à la précision de la finalité du changement. Si, dans le cas d'une action, il paraît admissible qu'une finalité précise existe (être une statue ou une maison), la potentialité étant fixée par l'agent, en revanche, dans le cas d'un changement naturel d'une substance, il paraît questionnable qu'un potentiel comporte en soi un état précis, comme le souligne Coope en prenant l'exemple d'une feuille poussée par le vent (*Ibid.*)[6] :

> *It is hard, then, to see what ground there is for thinking that this [such] movement is the actuality of a potential to be in some particular place, rather than another (...) Consider, for instance, a dead leaf that is blown across the street by the wind. Is it really plausible to suppose that its movement is the actualization of some specific potential it has for being on the other side of the street?*

En conclusion, que pouvons-nous retenir du traitement d'Aristote ? Essentiellement, une analyse métaphysique innovante – l'analyse dispositionnelle – rendant compte du caractère dynamique, fluant et causal du changement. Nous lui devons d'avoir initié une caractérisation du processus *se déroulant*, à savoir notre figure (1)[7]. Par contre, son traitement ne rend pas compte d'une dimension importante du changement, à savoir sa caractérisation temporelle et spatiale, par exemple, dans le cas du mouvement, le fait pour une substance d'être dans des états différents à des temps différents. Il va falloir attendre les XVII^{e} et XVIII^{e} siècles (Descartes, Galilée et Newton) pour que cette dimension soit mise en avant avec la théorie cinématique du mouvement.

3 Simples « changements de Cambridge » *vs* « changements réels » : une explication en terme de « processus du changement »

Dans cette section, nous enjambons quelques 2000 ans d'avancées en physique et métaphysique. La conception contemporaine du changement, héritière de la physique classique, fait dépendre le changement du temps. En des termes actuels, le changement se définit comme le fait qu'une substance puisse porter des propriétés contradictoires (*F* et *non F*) à des temps différents. Par exemple, un objet *O* est froid à un temps T_1 et chaud à un temps T_2[8]. Il s'agit là d'un changement impliquant une qualité de l'objet, considérée comme propriété intrinsèque. Dans le cas d'un mouvement, un objet *O* possède une certaine localisation à un temps T_1 et une autre à un temps T_2. La localisation correspond à une relation externe que l'objet entretient avec son environnement.

Au début du XX^{e} siècle, Bertrand Russell propose la conception suivante du mouvement continu, en tant qu'un changement de localisation spatiale : « *Motion consists merely in the occupation of different places at different times* » (Russell, 1903). Ainsi, un mouvement continu d'un objet *O* n'est *rien de plus* qu'une série de faits correspondant à l'occupation par *O* de différentes positions Pos_i à des instants successifs I_j : $<O,Pos_1,I_1>$, $<O,Pos_2,I_2>$, *etc.* Le terme « changement de Cambridge » a été proposé par Peter Geach (1968, p. 13) pour dénoter cette

*since they are directed at the same end (in this case, being a house) and are in the same stuff (in this case, the bricks and mortar) (Physics III.3).*Cette analyse est détaillée dans (Coope, 2014).

[6] Le même questionnement vaut pour les substances se mouvant par elles-mêmes vers leur lieu naturel : dans quelle mesure possèdent-elles ce 'potentiel à être dans leur lieu naturel' ? Comme on le sait, cette thèse au coeur de la physique d'Aristote a été balayée par les lois de la physique classique, notamment la loi de la gravitation exprimant que des corps s'attirent mutuellement.

[7] Ce point est confirmé par l'analyse de David Charles (2015, pp. 204-205) : *What are 'changes' in Aristotle's account ? As a distinctive type of continuant, they are, it seems, unfoldings or processes (changings), with different properties at different times. They can, as already noted, move through space and end before they should. So understood, his 'changes' are not changes, understood as events. Events do not move through space (...) To conclude: Aristotle's 'changes' (kinêseis), as continuants, are processes, not events.*

[8] Par souci de complétude, signalons qu'un autre type de changement peut être considéré comme lorsque des parties d'un même objet portent simultanément des propriétés contradictoires (par exemple, une partie est froide tandis qu'une autre est chaude). Dans cet article, nous nous intéressons plus particulièrement au changement global de propriété dans le temps.

conception du changement continu promue par des philosophes de Cambridge dont John McTaggart et Bertrand Russell[9].

Cette conception a été critiquée par plusieurs philosophes au début du 20ème siècle, au motif qu'elle ne rend pas compte du caractère *dynamique* du déplacement. Comme Henri Bergson (1946) a pu l'exprimer, le mouvement est traité « *comme si il était fait d'immobilités* »[10]. La question qui se pose est de savoir si, au-delà du sentiment intuitif que le « compte n'y est pas », il existe des arguments pour faire une place sur le plan ontologique à une entité correspondant à un « processus physique de transition », permettant de parler d'un « changement réel » (pour reprendre le terme de Geach (1968, p. 13)).

On peut tout d'abord noter que la métaphysique contemporaine donne crédit à l'existence de *Faits*, ceux-là mêmes auxquels il est fait référence dans les formulations de la théorie *at-at*. La thèse de l'existence des *Faits* a été défendue notamment par Kit Fine (1982) et David Armstrong (1997). Un *Fait* (ou « circonstance », selon la terminologie de Fine, ou encore « état d'affaires », pour reprendre le terme d'Armstrong), est une entité *complexe* constituée d'une substance (ex : 'Paul'), d'une propriété (ex : 'Être à côté de Marie') et d'un temps (ex : 'Maintenant') : <*Paul*, *Être à côté de Marie*, *Maintenant*>)[11]. L'existence simultanée à un instant donné d'une substance et d'une propriété ne signifie pas pour autant que la substance exemplifie la propriété à cet instant. Le *Fait* correspond à un lien interne unissant, à un instant donné, substance et propriété/relation en une entité à part entière. L'argument principal de l'existence des *Faits* est qu'ils constituent un *véri-facteur*, autrement dit ce qui rend vrai dans le monde des propositions comme 'Paul est à côté de Marie'. Les Faits auxquels il est fait référence dans la théorie *at-at* ont pour constituant une propriété de localisation spatiale d'une substance.

Ontologiquement parlant, on tient donc la brique de base d'une série de faits successifs. Par contre, comme le précise la philosophe Carol Cleland (1990), dont nous allons reprendre les propositions, quelle que soit la théorie retenue (y compris celle récente de Graham Priest (1985) défendant la thèse qu'un objet peut occuper plusieurs places à un même moment), aucune ne solutionne les limites de la théorie *at-at* à rendre compte de la dynamique du mouvement (*ibid.*, p. 264) :

> *Both the at-at theorist and Priest share a common assumption, viz., that moving can be completely analyzed in terms of the mere occupancy of places. As a consequence neither is able to provide us with an account of the transition in place ostensibly involved in the flight of an arrow. What is needed is an account of how it is possible for an arrow to get into and out of different places, as opposed to a description of which places just happen to be occupied at different times*

Pour rendre compte de la dynamique du mouvement, la démarche de Cleland, dans son (1990), consiste à réviser l'analyse dispositionnelle d'Aristote en la soumettant aux théories physiques contemporaines. Ainsi, pour rendre compte de la notion d'*effort*, au cœur de la physique Newtonienne, Cleland pose l'existence d'une entité correspondant à un « état actif de mouvement », qu'elle nomme « tendance opérante » (*Ibid.*, p. 266) :

> *In order to distinguish the sort of tendency which seems to be involved in « endeavouring » from the passive tendencies (or latent capacities) ordinarily associated with dispositional properties, I will frequently refer to "endeavourings" as "operative tendencies".*

L'existence d'une tendance opérante est justement ce qui distingue le fait, pour un objet, de passer dynamiquement à travers des états plutôt que d'être statiquement dans des états différents. Toutefois, selon Cleland, l'existence d'une tendance opérante, qu'elle nomme

[9] Cette conception porte également le nom de théorie « at-at ». Pour rendre compte du caractère continu du changement, la plupart des versions privilégient comme temps l'*instant* indivisible.

[10] Galton (2017) nous rappelle que le philosophe William James avait exprimé une opinion semblable dans son (1909, Lecture 6) : *Whatever motion really may be, it surely is not static; but the definition we gained [la théorie at-at] is of the absolutely static. It gives a set of one-to-one relations between space-points and time-points, which relations themselves are as fixed as the points are. It gives* positions *ad* infinitum*, but how the body gets from one position to another it omits to mention. The body gets there by moving, of course; but the conceived positions, however numerously multiplied, contain no element of movement (…)*.

[11] Pour des raisons de place, dans cet article, nous supposons sans justifications l'existence réelle, et même physique, du temps et de propriétés comme celles rendant compte d'une localisation spatiale.

également « processus causal de transformation », ne signifie pas pour autant qu'elle soit suivie d'effets. Ainsi, par exemple, l'existence d'une « tendance opérante à changer de place » reste insuffisante pour qu'un tel changement existe, si cette tendance se voit contrariée dans ses effets par l'existence concomitante d'autres tendances (*Ibid.*, p. 273) :

> *Indeed the failure of some of these tendencies to terminate in the changes towards which they are directed can readily be explained in terms of their lawful interactions with other tendencies; in the case of the globe, the outward centrifugal force is said to be exactly balanced by the inward centripetal force.*

L'exemple du globe, considéré ici, correspond à une expérience que chacun peut mener en faisant tourner autour de soi un globe (ou tout autre objet) maintenu par une ficelle. Selon Cleland, le fait que l'on puisse sentir la tension dans la ficelle, autrement dit le fait que la tendance de l'objet à être éjecté soit un observable mesurable, constitue un argument décisif pour son existence physique (*Ibid.*, 273) :

> *Given their crucial role in physical explanation and theory, I propose that we admit operative tendencies to be elsewhere into our ontology as primitive properties of physical objects. We can think of them as physicists think of instantaneous vector quantities, viz., as uneliminable proclivities of varying degrees of strength.*

Finalement, suivant Cleland, nous devenons capables de distinguer ontologiquement les « simples changements de Cambridge » des « changements réels ». Un *changement de Cambridge* est une série temporelle de faits correspondant, pour un objet, à des possessions successives de propriétés distinctes. Un *changement réel* est un changement de Cambridge occasionné, au sein de l'objet, par un processus de transformation de propriétés. Le fait de préciser « au sein de l'objet » est important en posant comme condition qu'un processus est actif de la part de l'objet et qu'il est la cause du changement de Cambridge. Cela permet d'exclure des situations où, par exemple dans le cas d'un mouvement, un objet est transporté (il se meut donc) sans pour autant que l'objet soit causalement et énergiquement responsable du mouvement. A contrario, des processus peuvent exister sans produire de changement, lorsque leurs effets se voient contrariés par ceux d'autres processus. Un tel engagement ontologique conduit finalement à distinguer le *processus du changement* du *changement* lui-même.

À titre d'illustration, considérons la situation d'un objet exerçant une pression sur un autre objet, par exemple ma main poussant une porte. Tant que la porte résiste à ma poussée, aucun mouvement n'existe. Pourtant, deux processus opposant leurs effets existent, à savoir : un processus de changement de place de ma main et, en réaction, un processus de changement de place de la porte[12]. À supposer que ma poussée s'intensifie en force, au point d'ébranler la porte. Nous considérons que cet ébranlement correspond à la naissance d'un nouveau processus – le processus de rotation de la porte sur ses gonds – venant s'ajouter aux deux processus déjà présents (ma poussée de la porte et sa réaction), lesquels continuent d'exister. Ce processus de mouvement conduit à l'ouverture de la porte.

Pour conclure, dressons le bilan de l'avancée que nous fait faire Cleland, par rapport à Aristote, dans l'analyse du changement (qualitatif). Nous avons conclu en Section 2 que Aristote avait contribué à rendre compte du changement *continu* plutôt que du changement *constaté*, en évoquant les deux figures ontologiques contemporaines que sont le *processus* et l'*événement*. Avec Cleland, nous précisons en quelque sorte le changement continu pour distinguer, d'un côté, le *processus du changement*, et de l'autre, le *changement* lui-même – ce dernier correspondant au changement Aristotélicien (restreint aux non actions)[13]. Dans la suite de l'article, nous allons confronter ce résultat aux dernières conceptions du processus présentées dans la littérature.

[12] Pour être complet, nous devrions également mentionner le processus des gonds de la porte correspondant à une force de frottement.

[13] Geach, dans sa caractérisation du « changement réel », l'identifie également à l'actualité Aritotélicienne (1968, p. 14) : *I suggest that when we have a narrative proposition corresponding to a 'real' change, there* is *an individual actuality – an imperfect actuality, Aristotle calls it – that is the change ; but not when a* mere *'Cambridge' change is reported*. Par contre, contrairement à Cleland, il ne va pas jusqu'à distinguer et reconnaître dans son inventaire ontologique le processus du changement.

4 Trois conceptions du processus : les raisons de notre choix

À l'instar de Cleland, plusieurs auteurs ont défendu ces dernières années une figure du processus comme *continuant dynamique* (Stout, 1997, 2003, 2016 ; Galton, 2006, 2008 ; Galton & Mizoguchi, 2009). Récemment, nous avons suivi cette voie (Kassel, 2017, 2018, 2019). Cette figure continue néanmoins d'être contestée par d'autres auteurs, s'arque boutant au contraire sur une vue du processus comme *occurrent* ne pouvant changer dans le temps (Smith, 2012) et, plus spécifiquement, comme constituant d'événements (Crowther, 2011, 2018)[14]. En Section 4.2, nous entendons contribuer aux débats en cours en proposant une caractérisation précise du processus physique sous forme de quatre propriétés essentielles.

Parallèlement, dans ses dernières publications, Galton a proposé une caractérisation des processus les situant en dehors de la strate physique pour en faire des entités abstraites mentales (Galton, 2012, 2018). En Section 4.3, nous analysons sa figure du processus abstrait et proposons de la retenir pour définir le changement lui-même.

En préambule, nous précisons en Section 4.1 quelques engagements ontologiques de base sur lesquels nous comptons nous appuyer dans nos argumentaires.

4.1 Nos engagements de base

Précisons en effet que nous poursuivons le projet de Peter Strawson (1959) d'établir une métaphysique descriptive visant à décrire « *la structure actuelle de notre pensée à propos du monde* ». L'objectif est ainsi d'établir des catégories et notions rendant compte de la façon dont nous concevons le monde et dont nous en parlons. Ces catégories peuvent être de sens commun ou savantes suivant les discours et théories les mobilisant. Par contre, nous considérons que leur identité ne peut être révisée au prétexte de viser à rendre compte de la réalité ultime du monde. En cela, nous nous dissocions notamment des philosophes des processus qui, prenant prétexte du fait que les sciences physiques nous enseignent que les objets matériels du niveau mésoscopique sont ultimement constitués, à un niveau nano, d'un tourbillon de particules et de vide, les identifient *réellement* à des masses dynamiques (voir, par exemple, (Seibt, 2008)).

Par ailleurs, adoptant une perspective contemporaine de l'ontologie, nous retenons comme mode principal de structuration du monde deux types de réalité : physique et mentale & sociale (voir par exemple (Poli, 2006) pour un exposé de ces deux « strates » et de leurs niveaux). Les termes « concret » et « abstrait » utilisés jusqu'à présent peuvent être identifiés à ces deux types de réalité. Ainsi, lorsque nous parlons d'un processus « concret », cela revient à le localiser dans la région spatio-temporelle du monde physique. En revanche, lorsque nous évoquons un événement ou un processus « abstrait », nous situons ces entités dans la strate mentale.

4.2 Le processus comme continuant dynamique

La thèse que nous entendons défendre quant à la nature du processus physique comporte un volet intensionnel et un volet extensionnel. En effet, les propriétés que nous attribuons aux processus dépendent étroitement de la classe des individus considérée.

En intension, nous caractérisons le *processus physique* comme quelque chose :

(p_i) existant pleinement à des instants ;

(p_ii) portant des propriétés à des instants ;

(p_iii) pouvant changer dans le temps ;

(p_iv) *énacté* par un objet physique.

[14] En réponse à Stout, la philosophe Helen Steward a également pris part récemment au débat en proposant une position médiane. Steward (2013, 2015) maintient que les processus sont des occurrents mais concède que ceux-ci peuvent changer dans le temps, à la manière de continuants. Il s'avère que la classe des processus considérée par Steward est beaucoup plus large que celle que nous allons retenir, ce qui explique sa position. Quoi qu'il en soit, pour des raisons de place dans cet article, nous ne pourrons pas exposer ses arguments.

En extension, des exemples sont : un processus de déplacement d'un objet physique ou de rotation de l'objet sur lui-même ; la croissance en taille d'un corps physique ; le mûrissement d'un fruit ; l'oxydation d'un objet métallique ferreux ; la fonte d'un glacier. À ces exemples s'ajoutent des processus ayant une cause intentionnelle : marcher, courir. Intuitivement, marcher et courir sont des espèces de processus de déplacement d'une personne.

En défense de la thèse, nous allons considérer tour à tour chacune des propriétés faisant partie de notre caractérisation des processus physiques. Nous allons notamment être amené à préciser nos engagements ontologiques sur la dualité physique-mental ainsi que sur le temps.

4.2.1 Les processus existent pleinement à des instants

Avant de prendre position, il convient de préciser en quel sens nous entendons l'expression « exister pleinement ». Cette même expression qualifie habituellement les objets physiques.

Kit Fine (2006) a apporté des précisions dans sa défense du 3-dimensionalisme des objets physiques[15]. Selon Fine, lorsque nous affirmons à propos des objets physiques qu'ils « existent pleinement à des temps » (ou qu'ils « existent dans leur entièreté à des temps »), cette expression fait référence à deux notions d'*existence* – respectivement dans le temps et dans l'espace – qu'il convient de distinguer. Dans l'ordre, l'objet *existe* dans le temps, cette existence n'admettant pas de degré (n'étant pas une question de « plus ou moins ») : on peut penser ici à une existence de l'objet dans sa pleine *identité* (au sens d'*essence*). Par ailleurs, du fait qu'il existe, l'objet matériel est *étendu* dans l'espace et, cette fois, l'expression « pleinement » (ou « dans son entièreté ») traduit le fait qu'il *occupe complètement* une région spatiale (tout en occupant partiellement chaque partie de cette région).

Dans l'énoncé de notre propriété (p_i), il convient d'entendre l'expression « existant pleinement » au sens d'une existence dans sa pleine *identité*. De ce fait, nous rapprochons les processus des objets en les assimilant à des continuants 3-D. Ceci pose la question de l'extension temporelle des processus et, corrélativement, de leur existence à des *intervalles* de temps – alors que nous affirmons qu'ils existent à des *instants*.

Rappelons à ce sujet que, dans le domaine de l'ontologie appliquée, l'extension temporelle des processus et la théorie du perdurantisme pour expliquer la persistance des processus sont des principes gravés dans le marbre. Dans l'ontologie BFO (Smith, 2012), le processus est identifié à une entité 4D occurrente. Pour l'ontologie DOLCE, ces principes, déjà solidement ancrés dès la version initiale (Masolo *et al.*, 2003), ont été repris récemment par Nicola Guarino (2017), qui a même proposé une analyse plus détaillée des parties temporelles des processus et des événements. Pour Galton, enfin, il ne fait aucun doute qu'un processus est étendu temporellement et que, corrélativement, le processus existe à un intervalle de temps[16].

La défense du 3-dimensionalisme des processus que nous opposons se fonde avant tout sur la classe des processus considérée. Rappelons que, pour nos processus, nous avons retenu la notion de *processus de transformation de propriété* de Cleland (1990). Elle se fonde également sur le fait de considérer des événements abstraits.

Pour l'ensemble des auteurs précisant actuellement la catégorie de processus physique celle-ci recouvre des processus tels 'Ecrire une lettre', 'Remplir un formulaire' ou 'Donner une conférence' qui, à l'évidence, ne correspondent pas à de simples « processus de transformation

[15] Le 3-dimensionalisme s'oppose à la thèse du 4-dimensionalisme selon laquelle les objets sont étendus à la fois spatialement et temporellement, et possèdent de fait des parties temporelles. Une *même* personne (selon la vue 3D) existant à des temps différents T_1 et T_2 est considérée selon la vue 4D comme *deux* personnes *différente*s à ces temps T_1 et T_2. Plus exactement, selon la vue 4D, la personne au temps T_1 et la personne au temps T_2 sont considérées comme étant des parties temporelles distinctes d'une même et unique personne. En conséquence, un objet 4D ne peut changer qualitativement, selon la conception du changement (de sens commun) que nous avons adoptée dans l'article. Le lecteur intéressé trouvera une défense du 4-dimensionalisme dans (Sider, 2001).

[16] Selon Galton, les processus n'existent pas à des *instants* – comme l'exprime notre propriété (p_i) – mais à des *intervalles* de temps, et ceci vient contredire la théorie 'présentiste' du temps selon laquelle seuls des *instants-présents* existent (Galton, 2017, p. 167) : *This idea* [selon laquelle les processus sont des particuliers concrets] *raises problems for the traditional instant-based model of time, since processes, being inherently temporally extended, can only exist over intervals, not at instants*.

de propriété ». Une analyse courante de ces phénomènes d'agence *téliques* (comportant en soi une fin) revient à identifier, d'une part, un événement comme occurrence complète (ex : le fait accompli d'écriture d'une lettre) et, d'autre part, un processus le constituant. Une telle analyse, comme nous l'avons rappelé en Section 2 en nous référant à Coope (2009), est clairement héritée de l'analyse Aristotélicienne des processus, avant d'avoir été renforcée par la thèse de la constitution d'événements par des processus (Mourelatos, 1978). Cet héritage, consistant à considérer des processus téliques ayant vocation à s'achever en un événement complet, est notamment assumé par Charles (2018). L'argument retenu pour justifier l'existence de tels processus est de nature linguistique. Comme on peut le voir chez Stout (2016), les processus sont définis comme étant ces entités occurrentes pour lesquelles on peut dire qu'elles étaient/sont/seront en train d'*occurrer* (*avoir lieu*, *se passer*) : *[Processes] are things that were, are or will be happening – like someone reading or my writing this paper for instance*.

Dans Kassel (2017, 2019), nous avons proposé un cadre ontologique étoffé (comportant, à côté des processus physiques, des *événements* abstraits et *des faits physiques* temporaires) permettant de dénier l'existence de tels 'processus d'événements'. Dans cet article, pour des raisons de place, nous nous contentons de résumer notre analyse ontologique. En premier lieu, signalons que, pour la notion d'*événement*, nous retenons un événement *abstrait* correspondant à un construit humain. Concernant l'analyse des phénomènes d'agence comme 'Écrire une lettre', nous identifions bien un événement, par contre notre conception non télique des processus nous interdit de considérer une espèce de 'macro' processus. À la place, nous considérons que la *réalisation* de phénomènes d'agence tel 'Écrire une lettre' donne lieu à de nombreux processus physiques (ex : des gestes d'écritures, des processus impliquant un crayon). L'événement n'est pas *constitué* de ces processus mais *est réalisé par* eux, lorsqu'il occure. De fait, l'argument linguistique évoqué par Stout (cf. supra) devient un argument en faveur de l'existence d'événements. Nous affirmons que les entités occurrentes pour lesquelles on peut dire qu'elles étaient/sont/seront en train d'*occurrer* (*avoir lieu*, *se passer*) sont des événements, et non des processus. Il s'agit d'événements en train d'être réalisés, comme pour : « Je suis en train d'écrire une lettre », « j'étais en train de traverser la rue », « je serai en train de partir en vacances », ou encore « une bagarre entre deux hommes à l'extérieur du nightclub est en cours »[17].

En restreignant ainsi notre classe de processus, par rapport aux théories courantes, non seulement nous défendons le 3-dimensionalisme de nos processus mais encore nous défendons qu'ils puissent exister à des instants et non à des intervalles de temps.

4.2.2 Les processus portent des propriétés à des instants

En retenant comme notion d'existence celle d'une *identité*/*essence* endurant dans le temps, nous admettons que les processus portent des propriétés à des temps – notre propriété (p_ii), particularisée aux instants. Certaines propriétés sont essentielles en constituant l'identité du processus, tandis que d'autres sont contingentes et correspondent à des manières d'être temporaires des processus.

La phénoménologie des processus nous indique que ceux-ci peuvent être, à des instants : *rapides*/*lents*, *bruyants*/*silencieux*, *chaotiques*/*réguliers*. On peut noter qu'ils ont des manières d'être distinctes de celles des objets (Hacker, 1982) : un processus n'a pas de couleur, de masse ou de volume, en revanche il se caractérise par sa vitesse, sa direction, sa sonorité, son amplitude. Ce constat justifie du reste que les objets et les processus soient considérés comme appartenant à deux classes distinctes de continuants.

On notera également que les processus entretiennent temporairement des relations avec d'autres processus. Reprenons notre exemple d'une situation où un objet, en exerçant une poussée sur un autre objet, le fait se mouvoir. Selon notre analyse, dans une telle situation, un

[17] Dans cette dernière phrase, nous considérons que le terme 'bagarre entre deux hommes' se réfère à un événement. Ce même exemple est traité par Stout au moyen de la phrase : « This is a process of two men fighting outside the nightclub ». On notera que la référence au processus contraint à recourir à un type de phrase ne relevant pas du langage courant.

processus de mouvement du premier objet 'perpétue' un processus de mouvement du second objet, cette relation correspondant à une propagation de causalité.

Dans la littérature récente, deux auteurs – Barry Smith (2012) et Thomas Crowther (2011, 2018) – dénient aux processus toute possibilité de porter des propriétés à des instants. Leur thèse (non p_ii) est fondée sur le fait d'assimiler les processus à des occurrents 4D. Au-delà, chacun des chercheurs met en avant des arguments différents. Pour illustrer les arguments convoqués, considérons la situation suivante : *Marie marche à la vitesse de 4km/h à l'instant T*. Rappelons que, selon notre traitement, un processus particulier de marche est actif et ce processus porte la propriété '*A pour vitesse 4km/h*' à l'instant *T*.

De son côté, Smith préconise dans BFO de considérer la vitesse comme une propriété de l'objet se mouvant (2012, p. 479)[18] : *Note that we could view speed in BFO terms as a (non-rigid) quality of the moving object, a view conformant with our way of speaking when we talk, for example, of the speed of light, or the speed of the earth, or the speed of a billiard ball*. Le problème que nous voyons est le manque de sémantique de cette propriété. Lorsque nous parlons de la vitesse d'une boule de billard, il est implicite qu'il s'agit de sa vitesse de déplacement. Mais, à supposer qu'un objet participe simultanément à plusieurs processus, comment faire pour savoir à quel processus rattacher la propriété (sauf à considérer une propriété comme 'vitesse de marche', ce que Smith s'interdit) ? Ce problème a été soulevé par Galton et Mizoguchi (2009, p. 79) :

> *To say that Mary is slow at a particular time is meaningless unless we specify in what respect she is slow; this could be any range of activities such as walking, speaking, thinking, etc. and since one such activity must be specified in ascribing slowness to Mary it is clear that it is the activity rather than Mary herself that is described as slow.*

Pour Crowther (2018), plutôt que de considérer qu'un processus de marche porte la propriété '*A pour vitesse 4km/h*' au temps *T*, il convient de considérer que l'objet se mouvant – Marie – porte la propriété '*Marcher à la vitesse de 4km/h*', toujours au temps *T*. On notera qu'un tel traitement n'est pas opposé au nôtre, à condition d'admettre les deux propriétés. La seconde propriété est pour nous une propriété relationnelle mettant en relation *Marie* et un processus particulier de marche, ce dernier ayant une propriété de vitesse. Mais, si cette dernière attribution est interdite, comment rendre compte de la sémantique de la propriété '*Marcher à la vitesse de 4km/h*' ? Crowther peut identifier la propriété '*Marcher*' à un type de processus, mais quel lien faire avec une vitesse, si le processus est assimilé à un occurrent étendu ? Nous estimons que le problème est repoussé, plutôt que résolu.

4.2.3 Les processus peuvent changer dans le temps

Venons-en à notre propriété (p_iii), à savoir la possibilité pour les processus de changer en portant des propriétés contraires à des temps différents. Cette thèse a notamment été soutenue par Galton (2006, p. 6) :

> *Like objects, processes can change: the walking can get faster, or change direction, or become limping. All around us processes undergo changes: the rattling in the car becomes louder, or change rythm, or may stop, only to start again later. The flow of the river becomes turbulent; the wind veers to the north-west.*

Considérons, dans la lignée des exemples que nous venons de prendre, un épisode de marche d'une personne et le fait qu'au cours de cet épisode cette personne à un instant « hâte le pas », autrement dit accélère. Notre analyse ontologique est la suivante : à chaque instant de l'épisode de marche, un processus de marche est pleinement présent (p_i) ; ce processus porte à chaque instant une vitesse (p_ii) ; à l'instant où la personne hâte le pas, la vitesse du processus de marche change de magnitude. C'est donc bien le processus qui a changé.

Cette propriété (p_iii) découle directement de (p_i) et (p_ii). Pour renforcer cette thèse, nous ajouterons le constat selon lequel nous pouvons agir sélectivement sur des processus de sorte à les modifier, voire à les bloquer. Par exemple, nous pouvons agir sur un fruit pour (1) ralentir

[18] Par ailleurs, Smith recommande de considérer le « profil de processus » universel '*Marcher à vitesse de 4 km/h*' dont le processus particulier de marche de Marie est une instance. Mais ceci est sans conséquence sur notre critique.

ou au contraire accélérer son mûrissement, mais également, et le cas échéant, pour (2) stopper sa chute. D'une façon générale, nous avons l'habitude, dans nos actions courantes, de moduler nos efforts pour faire varier proportionnellement des processus, par exemple lorsque nous poussons avec plus ou moins d'intensité des objets. Dans une situation de conduite automobile, en appuyant plus ou moins sur la pédale d'accélérateur nous faisons varier la vitesse de notre voiture.

Il est inutile de reprendre ici dans le détail les critiques formulées par les tenants de la vision 'occurrent 4D' des processus. Leur argument principal revient à dire que, comme un processus est déjà un changement, il ne peut lui-même changer. De notre côté, a contrario, c'est bien le fait de distinguer le 'processus du changement' du 'changement' lui-même qui nous permet d'accréditer la thèse du changement du processus. Une question reste toutefois en suspend qui est de savoir si le changement de processus peut être rapproché du changement d'objet. La réponse que nous apportons, suivant toujours (p_i) et (p-ii), est qu'il s'agit dans les deux cas d'un changement de propriétés dans le temps.

4.2.4 Les processus sont *énactés* par des objets physiques

Pour finaliser la caractérisation de nos processus physiques, évoquons un dernier engagement ontologique, à savoir le fait qu'un processus ne soit pas un continuant flottant dans l'air, mais soit ancré dans un objet support (notre propriété (p_iv)) : il s'agit du mouvement d'une *flèche*, du mûrissement d'un *fruit*, de la fonte d'un *glacier*, etc. Pour rendre compte de ce lien fort constitutif, nous reprenons à notre compte la relation d'*énaction* introduite par Galton et Mizoguchi (2009). Pour ces auteurs, dire qu'un objet « énacte » un processus revient à dire qu'un objet porte un processus « externe » ou « comportement » (*Ibid.*, p. 94) :

> *The key notion is that an object, considered from a particular point of view, is characterized in terms of the processes it enacts. These are what we call the external processes or behavior of the object. This behavior arises as a result of various internal processes which causally contribute to it.*

Cette caractérisation de la relation d'*énaction* par Galton et Mizoguchi repose sur une conception de l'objet comme interface entre des processus internes et externes. Cette conception, à son tour, met en avant une double hiérarchie de relations de constitution (et donc de dépendances existentielles) entre processus et objets situés à différents niveaux de la strate physique : un processus de marche d'une personne n'est possible que si des processus physiologiques énactés par les organes de la personne existent concomitamment ; ces derniers ne sont possibles que si des processus énactés par des tissus, des cellules, des molécules, etc. existent concomitamment[19].

Une telle conception de la relation entre processus et objets conduit à positionner au même niveau (en termes de priorité) ces deux primitives pour rendre compte de la réalité physique. Elle se trouve de facto en porte à faux avec la métaphysique traditionnelle des processus, laquelle donne la priorité au processus, arguant que les objets matériels « *are ultimately comprised of energy that is in an ongoing state of flux and motion* » (Rescher, 1996, p. 28). Plus particulièrement, la propriété (p_iv) se trouve en porte à faux avec une thèse courante en métaphysique des processus selon laquelle certains processus qualifiés d''impropres' n'ont pas de support physique (ne sont énactés par aucun objet), comme on le voit exprimé par Nicolas Rescher (*Ibid.*, p. 42) :

> *The distinction between 'owned' and 'unowned' processes also plays an important role in process philosophy. Owned processes are those that represent the activity of agents: the chirping of birds, the flowering of a bush, the rotting of a fallen tree. Such processes are ownership attributable with respect to "substantial" items. Unowned processes, by contrast, are free floating, as it were, and do not represent the activity of actual (i.e., more than nominal) agents: the cooling of the temperature, the change in climate, the flashing of lightning, the fluctuation of a magnetic field.*

[19] Dans son (2000), Peter Simons défend une thèse de l'endurance des objets physiques mettant en scène cette même double hiérarchie de relations de constitution entre des objets et des 'occurrences', ces dernières s'avérant correspondre à nos processus.

Concernant la question de la priorité accordée aux primitives d'*objet* et de *processus*, rappelons (cf. Section 4.1) que nous cherchons à établir une ontologie descriptive du monde. De ce fait, nous considérons que la connaissance scientifique que nous possédons de la constitution ultime des objets physiques (par exemple des artefacts mésoscopiques nous environnant) n'interfère en rien dans les concepts de sens commun de ces objets : une table demeure un objet physique solide.

Concernant la dépendance de processus vis-à-vis d'objets, à ce stade de la caractérisation de nos objets et processus, rien ne nous permet d'accréditer l'existence de processus 'impropres', bien au contraire. L'approche descriptive qui est la nôtre nous fait admettre l'existence de processus comme le refroidissement d'un lac ou le réchauffement climatique. Simplement, les entités support de ces processus sont des objets de la strate physique dépassant la taille mésoscopique pour relever de tailles macro, voire astro, Par ailleurs, il convient de noter que la caractérisation de l'objet physique en ontologie formelle est plutôt celle d'objets matériels solides. L'ontologie d'entités tels une rivière, une vague, un feu, un champ magnétique, reste largement à préciser. En avançant la propriété (p_iv), nous formulons l'hypothèse que de telles avancées nous conduiront à élucider des processus énactés par ces entités.

Dans cette section, nous venons de défendre la figure du processus comme continuant dynamique concret en réfutant celle du processus comme occurrent constituant d'événements. Ce faisant, pour revenir à notre fil conducteur du changement, nous avons identifié le processus à un « moteur » du changement ou 'processus du changement'. La figure du *processus abstrait* défendue par Galton (2012, 2018) va nous donner l'occasion d'accorder au changement lui-même un statut ontologique.

4.3 Le processus comme pattern abstrait d'occurrence

Depuis 2012, Galton promeut une figure ontologique différente du processus. Nous pouvons même dire « très différente »[20] puisque, d'une entité *concrète*, Galton en fait une entité *abstraite* (2012, *Abstract*) : *We regard processes as abstract patterns of behaviour which may be realised in concrete form as actually occuring states and events*. Pour appréhender cette nouvelle figure, voyons en quel sens Galton entend les termes « abstrait », « patron de comportement », « réalisation », « état » et « événement ».

Dans son (2012), Galton prend soin de préciser deux notions, celles d'*abstraction* et de *réalisation*. Notons tout d'abord que, pour Galton, états et événements sont respectivement des continuants et des occurrents concrets (Galton continue de souscrire à la notion Davidsonienne d'événement). Plus précisément, pour Galton, un *état* est caractérisé comme un comportement *continuable* (ex : marcher) tandis qu'un *événement* est un comportement *répétable* (ex : marcher jusqu'à la gare)[21]. Dès lors, un *processus* est un type d'état/événement, la relation type-instance correspondant à une *réalisation* (*Ibid.*, §6 *Realisations*) : *Such instantiations are concrete realisations of these continuable or repeatable behaviours. As such they are fully determinate with respect to their spatial, temporal, and indeed all other characteristics*. Le processus est ainsi une entité abstraite :

> *The category of process is neither subordinate to nor superordinate to the categories of state and event; nor is it on the same footing as them with some immediate common superordinate category. On the contrary, processes belong in a completely different realm, the realm of abstract entities, patterns if you will, quite separate from the realm of spatio-temporal entities which includes both states and events*.

[20] Pour Galton, l'intention n'est pas de compléter le cadre précédemment établi dans ses articles de 2006, 2008 et 2009, mais bien de le réviser radicalement. Ainsi écrit-il (2012, p. 35) : *In this paper, I advocate a point of view which is in some respects utterly at variance from those expressed by the authors represented in Figure 1* [Allen, Moens & Steedman, Mourelatos, Pustejovsky, Sowa] – *and indeed from my own previous publications on this subject*.

[21] La différence est qu'un comportement comme 'marcher' peut être *poursuivi* dans l'instant (cette poursuite étant toutefois conditionnée à une décision de mettre fin au comportement ou à un état physique permettant cette poursuite), tandis que le comportement 'marcher jusqu'à la gare' ne peut être poursuivi, une fois arrivé à la gare, mais peut être *répété* quotidiennement.

Il reste à comprendre ce que Galton entend par "patron de comportement". La définition de cette notion fait l'objet de son (2018). En introduction, à titre d'exemple, Galton considère le processus de marche (*Ibid.*, p. 41) :

> *The process of walking, for instance, may be characterized by a particular pattern of movement, alternating forward swings of the legs resulting in an overall forward movement of the body. An actual realization of this pattern, viewed synoptically, is an event which consists of someone's starting to walk (that is, to realize the walking pattern), walking for a while, and then stopping. Viewed experientially, from moment to moment, we see a succession of instantaneous states, each of which may be characterized as a walking state, that is, a state in which the disposition and state of motion of the body parts is characteristic of one phase of the walking pattern.*

Cette caractérisation du processus de marche appelle deux remarques. Tout d'abord, l'expression « patron de mouvement » (spécialisée, à l'occasion, en « patron de marche ») utilise le terme « mouvement », sans que cette notion soit définie. Plus précisément, Galton ne ménage pas de place au mouvement dans son inventaire ontologique. Par ailleurs, une nouvelle catégorie d'*état* est introduite. Nous avions (dans son (2012)) l'*état de comportement* en tant que *comportement continuable*. Désormais, nous avons une succession d'*états instantanés*, chacun d'entre eux étant « caractérisé comme un état de marche », autrement dit comme un *état de comportement*. On voit ici un problème d'incohérence ontologique : comment l'*état de march*e, qui est un particulier concret, peut-il « caractériser » (dans une relation ne pouvant être que type-instance) des *états instantanés* ?

Reprenons, pour notre part, l'analyse du mouvement là où nous l'avons laissée en fin de Section 3. Nous allons voir qu'en cherchant à rendre compte complètement du mouvement, nous arrivons à faire une place dans notre inventaire ontologique à une entité proche du *patron abstrait d'occurrence* de Galton.

Pour caractériser le changement/mouvement continu, nous avons mis en avant (1) le processus du mouvement (notre processus physique), lequel (2) se manifeste par une série de faits instantanés (ces faits instantanés successifs résultent du processus). Nous avons fait une place dans notre inventaire ontologique aux faits instantanés (rappelons-le, en mobilisant les notions proches de *circonstance* de Fine (1982) et d'*état d'affaires* d'Armstrong (1997)). Par contre, nous avons laissé de côté la série – en tant que telle – de faits instantanés, pour la raison que nous n'acceptons pas dans notre inventaire ontologique d'entités existant sur un intervalle de temps. De fait, nous avons laissé de côté le mouvement. Comment en rendre compte ?

Une façon simple de traiter le mouvement (continu) est de faire appel à notre capacité de perception. Notons tout d'abord que, sous réserve que le mouvement ne soit pas trop rapide ou trop lent pour que nous l'identifiions, nous percevons les faits instantanés successifs de localisation de l'objet se mouvant. Dans le cas d'un mouvement global de l'objet (lorsque toutes ses parties changent de localisation spatiale), la série successive de faits n'est, ni plus ni moins, qu'une trajectoire dans une région spatio-temporelle. Or, et c'est une seconde donnée de la perception que nous mobilisons, cette trajectoire est quelque chose que nous percevons globalement, que nous unifions, et à laquelle nous attribuons une forme. Il paraît dès lors naturel d'assimiler le mouvement à cette trajectoire et de parler de la forme du mouvement, ce qu'attestent des expressions comme « mouvement *rectiligne*/*circulaire*/*oscillatoire* »[22].

Cette proposition nous paraît être en bonne cohérence avec la figure du processus abstrait comme patron d'occurrence défendue par Galton, tout en répondant aux deux remarques que nous formulions supra. D'une part, nous caractérisons ontologiquement le mouvement et, plus largement, le changement continu. Sur ce point on notera que, suivant la nature du processus et de sa manifestation, les faits instantanés peuvent varier et être, par exemple, des faits de sonorité distincts, mobilisant une modalité de perception différente, en l'occurrence l'ouïe[23]. D'autre

[22] Pour conforter cette indentification du mouvement à une conceptualisation d'une trajectoire, nous pouvons citer, en sémantique cognitive, les travaux du linguiste Ronald Langacker (1987) portant sur la caractérisation de verbes de 'mouvement physique'. Selon Langacker, des verbes comme *aller*, *partir*, *grimper*, *rouler*, etc., ont pour sens une conceptualisation d'un élément mobile occupant successivement des positions différentes dans le temps.

[23] Cette remarque nous invite à positionner la mélodie sonore à côté du mouvement spatio-temporel. Nous gardons cette invitation pour des travaux futurs.

part, nous distinguons bien les deux catégories d'*état* confondues, ou en tout cas mal caractérisées, selon nous, par Galton. La première, l'*état de comportement*, correspond à notre *processus du changement*[24]. La seconde, l'*état instantané*, correspond à nos *faits instantanés* successifs.

En synthèse, nous proposons de positionner l'ensemble des entités ontologiques que nous retenons en les illustrant sur l'analyse d'un épisode de marche comme suit (le lecteur pourra ainsi comparer avec l'analyse de Galton rappelée supra) :

> Lors d'un épisode (continu) de marche d'une personne, un *processus de marche* énacté par cette personne est pleinement présent à tout instant que dure l'épisode. Si l'intention de la personne est de se rendre d'un point A à un point B, le processus est le moyen de réaliser cet *événement* – son déplacement de A à B. Sans intention particulière ni raison de penser à sa marche, aucun événement n'existe. A noter toutefois qu'un observateur peut prêter à la personne l'intention d'un déplacement, auquel cas le processus sera considéré comme le moyen mis en œuvre pour réaliser l'événement – pensé cette fois par l'observateur. Au cours de l'épisode de marche, le processus peut être irrégulier et changer, par exemple, de vitesse ou de direction. Un observateur, percevant le déplacement en cours de la personne, l'identifie à un *mouvement*. Il reconnait ainsi la forme caractéristique d'un mouvement de marche, caractérisée par la forme globale du déplacement mais aussi, et surtout, par la forme des mouvements des jambes et du torse de la personne. Ces mouvements étant régulièrement coordonnés et répétitifs se traduisent par un motif se répétant dans la forme du mouvement de marche.

5. Remarques concluantes et perspectives

Dans cet article, nous avons précisé la nature ontologique du processus physique et, chemin faisant, ayant pris comme fil conducteur l'analyse du changement et du mouvement (en tant qu'une espèce de changement), nous avons en quelque sorte poursuivi le projet d'Aristote de défense de la *réalité* du changement. Adoptant toutefois une conception contemporaine de l'ontologie en admettant deux types de réalité – physique et mentale (et sociale) – nous avons positionné le changement dans la réalité mentale comme une espèce d'événement. La Figure 1 synthétise graphiquement le cadre ontologique global auquel nous parvenons.

Ce cadre ontologique renouvelle très largement la distinction entre 'continuants' et 'occurrents', prégnante en ontologie formelle et largement adoptée en ontologie appliquée. Rappelons qu'il repose sur des positions encore minoritaires à la fois en ontologie des processus (où la figure du processus comme matière constituante d'événements continue de régner) et en ontologie des événements (où la thèse d'événements concrets reste prédominante). Il repose également, en philosophie du temps, sur une thèse présentiste (soutenant que seul le présent existe, sous forme d'instants) faisant l'objet de nombreux débats. Il repose enfin sur la thèse de l'existence de faits physiques donnant lieu également à de nombreux débats. Par là-même, nous soulignons que la défense de ce cadre nécessite encore des efforts importants.

Au-delà, nous identifions une perspective importante devant nous amener à compléter notre cadre ontologique. Comme nous l'avons rappelé, nous avons suivi comme fil conducteur le changement *qualitatif*. Une perspective à ce travail est de se pencher sur le changement *substantiel* correspondant, rappelons-le, à la création et/ou destruction de substances concrètes. La défense de la *réalité* du changement substantiel, à laquelle s'est attelé récemment le philosophe Edward Lowe (2006), nous conduit à penser que nous devrions élargir notre classe des processus physiques pour intégrer des processus de maintien de l'intégrité de substances, à l'instar du processus de vie pour un être humain.

[24] On notera à ce propos que Cleland, dans son (1990), utilise également le terme « état de changement » comme synonyme de « processus de transformation de propriété ».

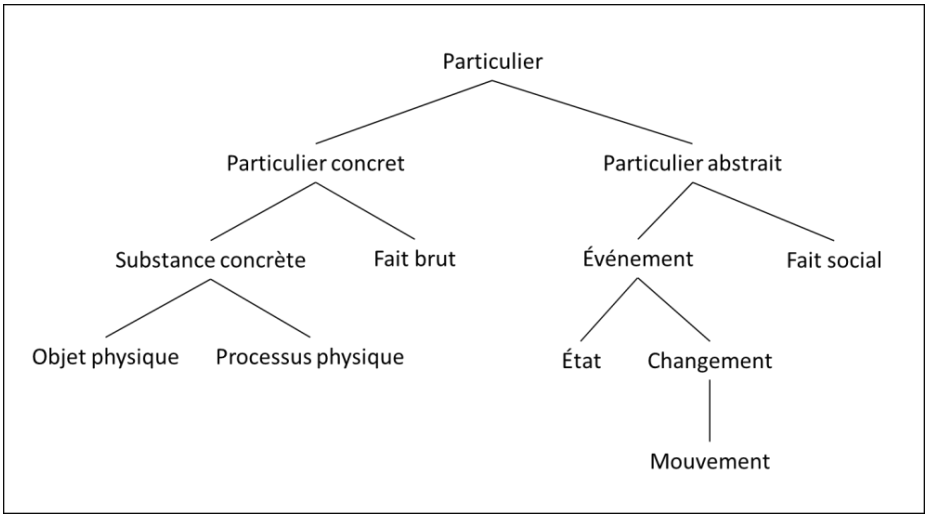


FIGURE 1 – Vue d'ensemble de notre cadre ontologique

Références

ANAGNOSTOPOULOS A. (2010). Change in Aristotle's Physics 3. *Oxford Studies in Ancient Philosophy*, 39, 33-79.

ARMSTRONG D.M. (1997). *A world of states of Affairs*. Cambridge University Press.

BARTHÉLEMY-SAINT-HILAIRE J. (1862). *Physique d'Aristote ou leçons sur les principes généraux de la nature*. Paris, Librairie philosophique de Ladrange.

BERGSON H. (1946). *The creative Mind*. New York, Philosophical Library.

CHARLES D. (2015). Aristotle's processes. In M. Leunissen (ed.), *Aristotle's Physics; A Critical Guide*, Cambridge University Press, pp. 186-205.

CHARLES D. (2018). Processes, Activities, and Actions. In R. STOUT (ed.), *Process, Action, and Experience*, Oxford University Press, pp. 20-40.

CLELAND C.E. (1990). The Difference Between Real Change and *Mere* Cambridge Change. *Philosophical Studies*, 60, 257-280.

COOPE U. (2004). Aristotle's Account of Agency in *Physics III.3*. In Proc. of *the Boston Area Colloquium in Ancient Philosophy*, XX, pp. 201-227.

COOPE U. (2009). Change and its relation to actuality of potentiality. In G. ANAGNOSTOPOULOS (ed.), *A compagnion to Aristotle*, Blackwell Publishing, pp. 277-291.

CROWTHER T. (2011). The Matter of Events. *The Review of Metaphysics*, 65(1), 3-39.

CROWTHER T. (2018). Processes as Continuants and Processes as Stuff. In R. STOUT (ed.), *Process, Action, and Experience*, Oxford University Press, pp. 58-81.

FINE K. (1982). First-Order Modal Theories III – Facts. *Synthèse*, (53), 43-122.

FINE K. (2006). In Defense of Three-Dimensionalism. *The Journal of Philosophy*, 3(12), 699-714.

GALTON A. (2006). On What Goes On: The ontology of processes and events. In R. FERRARIO & W. KUHN (eds.), *proceedings of the Fourth International Conference on Formal Ontology in Information Systems (FOIS2006)*, pages 4-11.

GALTON A. (2008). Experience and History: Processes and their Relation to Events. *Journal of Logic and Computation*, 18(3), 323-40.

GALTON A. (2012). The ontology of states, processes, and events. In M. OKADA & B. SMITH (eds.), *Interdisciplinary Ontology: Proceedings of the Fifth Interdisciplinary Ontology Meeting*, Open Research Centre for Logic and Formal Ontology, Keio University, Tokyo, Japan, pp. 35-45.

GALTON A. (2017). The Dynamic Present. In P. HASLE, P. BLACKBURN & P. OHRSTROM (eds.), *Logic and Philosophy of Time: Themes from Prior,* Aalborg University Press, pp. 167-187.

GALTON A. (2018). Processes as Patterns of Occurrence. In R. STOUT (ed.), *Process, Action, and Experience*, Oxford University Press, pp. 41-57.

GALTON A. & MIZOGUCHI R. (2009). The water falls but the waterfall does not fall: New perspectives on objects, processes and events. *Applied Ontology*, 4, 71-107.

GEACH P. & STOOTHOFF R. (1968). What Actually Exists. In *Proc. of the Aristotelian Society, Supplementary Volumes*, 42, pp. 7-30.

GUARINO N. (2017). On the semantics of ongoing and future occurrence identifiers. In H.C. MAYR G. GUIZZARDI, H. MA & O. PASTOR (eds.), *Conceptual Modeling, Proc. of 36th Int. Conf. ER 2017*, LNCS Springer, pp. 477-490.

HACKER P.M.S. (1982). Events and Object in Space and Time. *Mind*, 91, 1-19.

JAMES W. (1909). *A Pluralistic Universe: Hilbert Lectures at Manchester College on the Present Situation in Philosophy*. New York: Longmans, Green, and Co.

KASSEL G. (2017). Processus, événements et couplages temporels et causaux. *Revue d'Intelligence Artificielle*, 31(6), 649-679.

KASSEL G. (2018). Une alternative à la distinction 'continuant' vs 'occurrent'. In *Actes de la Conférence Nationale d'Intelligence Artificielle (CNIA 2018)*, pp. 127-136. https://afia.asso.fr/wp-content/uploads/2019/02/Ouvrage_web_promotion_AFIA_2018.pdf

KASSEL G. (2019). Processes Endure, Whereas Events Occur. In S. BORGO *et al.* (eds.), *Ontology Makes Sense: Essays in honor of Nicola Guarino*, IOS Press, Frontiers in Artificial Intelligence and Applications, vol. 316, pp. 177-193.

KOSMAN L.A. (1969). Aristotle's Definition of Motion. *Phronesis*, 14(1), 40-62.

LANGACKER R.W. (1987). Mouvement abstrait. *Langue française*, 76, 59-76.

LOWE E.J. (2006). How Real Is Substantial Change? *The Monist*, 89(3), 275-293.

MASOLO C., BORGO S., GANGEMI A., GUARINO N., OLTRAMARI A. & SCHNEIDER L. (2003). The WonderWeb Library of Foundational Ontologies and the DOLCE ontology. WonderWeb Deliverable D18, Final Report, vr. 1.0.

MOURELATOS A.P.D. (1978). Events, Processes, and States. *Linguistics and Philosophy*, 2(3), 415-434.

POLI R. (2006). Levels of Reality and the Psychological Stratum. *Revue internationale de philosophie*, 2(236), 163-180.

PRIEST G. (1985). Inconsistencies in Motion. *American Philosophical Quaterly*, (22), 339-346.

RESCHER N. (1996). *Process Metaphysics: an introduction to process philosophy*. Albany, Suny Press.

RUSSELL B. (1903). *Principles of Mathematics*. Cambridge, UK: Cambridge University Press.

SEIBT J. (2008). *Beyond Endurance and Perdurance: Recurrent Dynamics*. In C. KANZIAN (ed.), Persistence, Frankfurt: Ontos Verlag, pp. 133-164.

SIDER T. (2001). *Four-dimensionalism: An Ontology of Persistence and Time*. Oxford University Press, Oxford.

SIMONS P. (2000). Continuants and Occurrents. In *Proceedings of the Aristotelian Society*, Supplementary Volumes, 74, 59-75.

SMITH B. (2012). Classifying Processes: An Essay in Applied Ontology. *Ratio*, 25(4), 463-488.

STEWARD H. (2013). Processes, Continuants, and Individuals. *Mind*, 122(487), 781-812.

STEWARD H. (2015). What is a continuant? In *Proceedings of the Aristotelian Society*, Volume: LXXXIX, pp. 109-123.

STOUT R. (1997). Processes. *Philosophy*, 72(279), 19-27.

STOUT R. (2003). The life of a process. In G. DEBROCK (ed.), *Process Pragmatism: Essays on a Quiet Philosophical Revolution*, Rodopi, pages 145-57.

STOUT R. (2010). What are you causing in acting? In J.H. AGUILAR & A.A. BUCKAREFF (eds.), *Causing Human Action: New Perspectives on the Causal Theory of Action*, MIT Press, pp. 101-14.

STOUT R. (2016). The category of occurrent continuants. *Mind*, 125(497), 41-62.

STOUT R. (ed.)(2018). *Process, Action, and Experience*. Oxford University Press.

STRAWSON P. (1959). *Individuals. An Essay in Descriptive Metaphysics*. Methuen, London.

EpidNews : un explorateur de données épidémiologiques pour la surveillance des maladies animales

Alexis Delaforge[1,2], Rohan Goel[3], Samiha Fadloun[1], Sarah Valentin[4,5], Arnaud Sallaberry[1,2], Mathieu Roche[5], Pascal Poncelet[1]

[1] LIRMM, Université de Montpellier, CNRS, Montpellier, France

[2] Université Paul-Valéry Montpellier 3, Montpellier, France
alexis.delaforge@etu.univ-montp3.fr, arnaud.sallaberry@univ-montp3.fr

[3] BITS Pilani, Department of Computer Science Goa, India
f20140014@goa.bits-pilani.ac.in

[4] CIRAD, ASTRE, TETIS, Université de Montpellier, Montpellier, France
sarah.valentin@cirad.fr

[5] TETIS, AgroParisTech, CIRAD, CNRS, Irstea, Université de Montpellier, Montpellier, France
mathieu.roche@cirad.fr

Résumé : La détection et le suivi de foyers de maladies animales reposent sur l'analyse quotidienne de données épidémiologiques, généralement issues d'organismes officiels. Les médias digitaux, sources de données dites « non-officielles », sont utilisés de manière complémentaire par les systèmes de veille sanitaire en raison de leur réactivité et de la facilité d'accès à leur contenu. Cependant, l'extraction manuelle d'informations pertinentes à partir de ces données non-structurées est très couteuse en temps. Afin de simplifier cette tâche, nous proposons EpidNews (Goel *et al.* (2018)), un nouvel outil d'analyse visuelle qui permet d'explorer les données épidémiologiques multi-sources (officielles et non-officielles). EpidNews permet de visualiser la distribution et l'évolution spatio-temporelle des données grâce à une carte et un outil de sélection temporelle. Les données, qui représentent des foyers de maladie, peuvent être filtrées dynamiquement en fonction de leurs caractéristiques épidémiologiques (nom de la maladie, hôte, symptôme). Lors de leur affichage sur la carte, les données sont différenciées en fonction de leur source (officielle ou non-officielle) et de leur valeur. Des outils complémentaires adaptés à la représentation des données épidémiologiques sont proposés. Une carte de chaleur permet d'identifier les zones de faible et de forte densité d'occurrences de foyers. Un outil de sélection à main levée (lasso) permet de sélectionner un sous-ensemble de foyers et de visualiser leurs dates d'occurrence sur la barre temporelle. La distribution des données en fonction de leurs caractéristiques épidémiologiques est visualisée grâce à un graphique sunburst, qui joue également le rôle de filtre et peut être hiérarchisé en fonction des besoins de l'utilisateur. Les vues des différents outils se synchronisent automatiquement en fonction des filtres et des sélections dynamiques, facilitant l'interaction entre l'utilisateur, les filtres et la visualisation. EpidNews a été utilisé par un épidémiologiste afin de comparer les données officielles et non-officielles de foyers de peste porcine africaine sur une période de 21 mois. L'expert a pu visualiser la progression de la maladie vers l'Europe de l'Est, en distinguant les foyers de porcs domestiques et de sangliers. La comparaison des données officielles et non-officielles a permis de mettre en évidence de potentielles anomalies dans les données officielles (absence ou retard de déclarations). L'accès au contenu des données non-officielles à partir d'EpidNews a facilité l'évaluation de l'expert. Une vidéo de démonstration est disponible sur Youtube [1].

Mots-clés : Analyse visuelle, visualisation, épidémiologie animale, données multi-sources

Références

GOEL R., FADLOUN S., VALENTIN S., SALLABERRY A., ROCHE M. & PONCELET P. (2018). Epidnews : An epidemiological news explorer for monitoring animal diseases. In A. KERREN, K. KLEIN & Y. LI, Eds., *Proceedings of the 11th International Symposium on Visual Information Communication and Interaction, VINCI 2018, Växjö, Sweden, August 13-15, 2018*, p. 1–8 : ACM.

1. https://youtu.be/N8yfM42P4ME

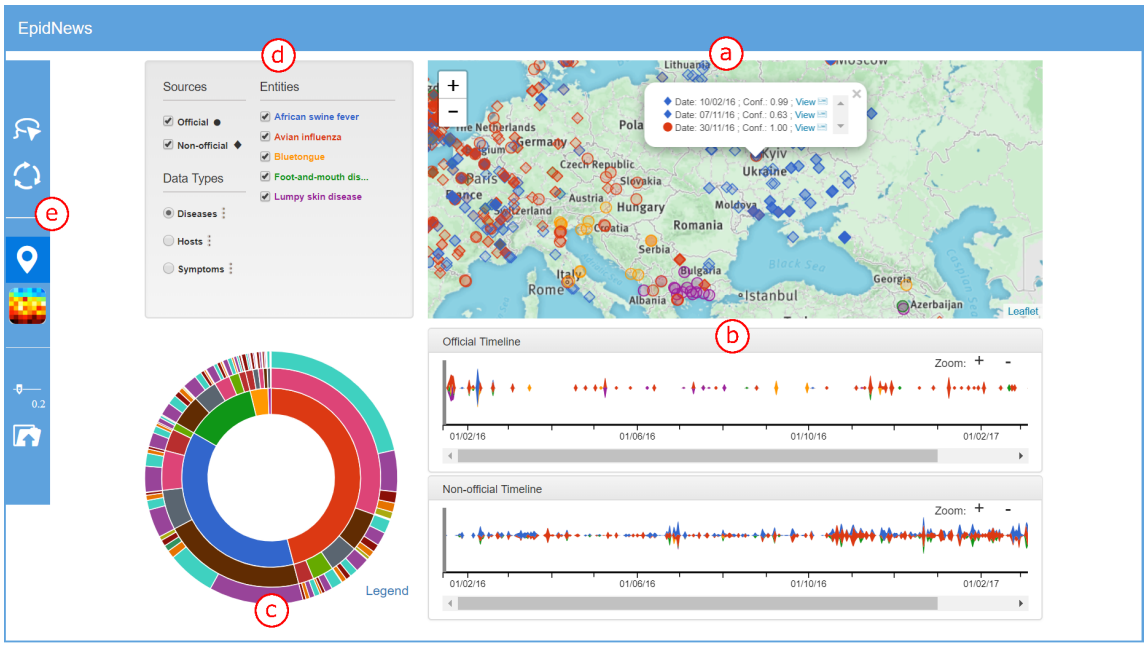


FIGURE *1 – Aperçu de l'outil EpidNews.* ***(a)*** *La carte montre l'emplacement des foyers de maladie à l'aide d'un cercle ou d'un losange, selon le type de source.* ***(b)*** *Les graphiques chronologiques permettent de comparer l'évolution du nombre de foyers issus des sources officielles et non officielles.* ***(c)*** *Le sunburst présente les relations entre les maladies, les hôtes et les symptômes dans une vue hiérarchique.* ***(d)*** *Le gestionnaire de données permet de manipuler les données représentées dans les autres vues (sources, type d'entité épidémiologique).* ***(e)*** *La barre d'outils offre d'autres fonctionnalités interactives.*

AgroLD: un graphe de connaissances pour la caractérisation des mécanismes moléculaires complexes impactant le phénome des plantes

Pierre Larmande[1,3], Gildas Tagny[2,3], Manuel Ruiz[2,3]

[1] UMR DIADE, IRD, Univ. of Montpellier, Montpellier, France.
pierre.larmande@ird.fr

[2] UMR AGAP, CIRAD, Montpellier, France
gildas.tagny_ngompe@cirad.fr, manuel.ruiz@cirad.fr

[3] SOUTH GREEN BIOINFORMATICS PLATFORM - Montpellier, France.

Résumé : La compréhension des interactions génotype-phénotype est un des axes les plus importants de la recherche en agronomie dont l'un des objectifs est d'accélérer la reproduction des caractères importants pour la production agricole. Or ces interactions sont complexes à identifier car elles s'expriment à différentes échelles moléculaires dans la plante et subissent de fortes influences de la part des facteurs environnementaux. Les technologies d'analyse haut-débit ne permettent de capturer que partiellement cette dynamique. Même si ces technologies sont de plus en plus performantes dans l'acquisition de données, notre connaissance du système reste encore parcellaire pour pouvoir comprendre les relations complexes existant entre les différents éléments moléculaires responsables de l'expression du phénome -ensemble des phénotypes observés pour un individu- . Cet objectif ne peut être atteint qu'en intégrant des informations de différents niveaux dans un modèle intégrateur utilisant une approche systémique afin de comprendre le fonctionnement réel d'un système biologique. Aujourd'hui, le Web sémantique propose des technologies pour l'intégration de données hétérogènes et leur transformation en connaissances explicites grâce aux ontologies.

Nous avons développé AgroLD (Venkatesan *et al.*, 2018) (Agronomic Linked Data - www.agrold.org), une base de connaissances reposant sur les technologies du Web sémantique et exploitant des ontologies du domaine biologique, afin d'intégrer des données issues de plusieurs espèces de plantes présentant un intérêt important pour la communauté scientifique, comme par exemple le riz, le blé et arabidopsis. Nous présentons les résultats du projet, qui portait initialement sur la génomique, la protéomique et la phénomique. AgroLD est aujourd'hui une base de plus de 100 millions de triplets créée à partir de plus de 50 jeux de données provenant d'une dizaine de sources de données, telles que Gramene (Tello-Ruiz *et al.*, 2018) et TropGeneDB (Hamelin *et al.*, 2012). Par ailleurs, nous avons utilisé une dizaine d'ontologies du domaine biologique, telles que Gene Ontology (The Gene Ontology Consortium, 2014) et Plant Ontology (Plant & Consortium, 2002) pour annoter et intégrer ces ressources. Pour cette phase, chaque jeu de données a été transformé à partir de sources sélectionnées et annotées sémantiquement en réutilisant les champs textuels correspondant avec des termes d'ontologies lorsqu'ils ont été fournis par la source d'origine. De plus, nous avons utilisé les services Web d'AgroPortal (Jonquet *et al.*, 2018) pour annoter sémantiquement des éléments supplémentaires tels que par exemple, les URIs correspondant à la taxonomie des espèces ou des éléments d'anatomie. Dans ces cas, nous avons généré des propriétés supplémentaires à partir des ontologies correspondantes, ajoutant ainsi 22% de triplets supplémentaires qui ont été validés manuellement.

L'objectif d'AgroLD est d'offrir une plate-forme de connaissances spécifiques du domaine agronomique afin de répondre à des questions biologiques complexes. De telles questions peuvent concerner le rôle de gènes spécifiques dans les mécanismes de résistance aux maladies des plantes ou de caractères de production identifiés à partir des analyses GWAS. Afin de rendre AgroLD accessible par un plus grand nombre d'utilisateurs, nous avons également développé une application Web proposant plusieurs interfaces de requêtes. Tout d'abord une interface simple qui permet aux utilisateurs de rechercher par mots-clés sur l'ensemble des valeurs de la base et ainsi de parcourir le contenu d'AgroLD. Puis une interface de recherche avancée qui permet de combiner du texte libre et des filtres à facettes ainsi que des services Web externes proposant ainsi une interface d'agrégation de données distribuées. AgroLD possède également une interface de visualisation des graphes qu'il est possible de configurer pour mettre en valeur certains types de relations. Finalement, un éditeur SPARQL propose un environnement interactif pour formuler des requêtes et manipuler des résultats. Actuellement, de nouveaux jeux de données sont en cours d'intégration. Ils portent sur les réseaux d'interaction protéine-protéine, les facteurs de transcription et réseaux de co-expression afin d'étendre les connaissances sur les mécanismes moléculaires. De nombreux développements sont également réalisés au niveau des interfaces de requêtes, notamment au niveau de la visualisation des graphes afin de fournir des outils plus dynamiques, interactifs et contextualisés. Enfin, une attention particulière est portée sur la qualité des données intégrées. Des méthodes de liage et de machine learning sont développées pour rechercher des liens et des ressources similaires dans la base de connaissances ou dans des ressources externes.
Mots-clés : Base de connaissances, Web sémantique, Agronomie, Génomique fonctionnelle, Phénome

IC 2019

Références

HAMELIN C., SEMPERE G., JOUFFE V. & RUIZ M. (2012). TropGeneDB, the multi-tropical crop information system updated and extended. *Nucleic acids research*, p. gks1105.

JONQUET C., TOULET A., ARNAUD E., AUBIN S., DZALÉ YEUMO E., EMONET V., GRAYBEAL J., LAPORTE M.-A., MUSEN M. A., PESCE V. & LARMANDE P. (2018). AgroPortal : A vocabulary and ontology repository for agronomy. *Computers and Electronics in Agriculture*, **144**, 126–143.

PLANT T. & CONSORTIUM O. (2002). The Plant Ontology Consortium and plant ontologies. *Comparative and functional genomics*, **3**(2), 137–42. Citation Key : Plant2002.

TELLO-RUIZ M. K., NAITHANI S., STEIN J. C., GUPTA P., CAMPBELL M., OLSON A., WEI S., PREECE J., GENIZA M. J., JIAO Y., LEE Y. K., WANG B., MULVANEY J., CHOUGULE K., ELSER J., AL-BADER N., KUMARI S., THOMASON J., KUMAR V., BOLSER D. M., NAAMATI G., TAPANARI E., FONSECA N., HUERTA L., IQBAL H., KEAYS M., MUNOZ-POMER FUENTES A., TANG A., FABREGAT A., D'EUSTACHIO P., WEISER J., STEIN L. D., PETRYSZAK R., PAPATHEODOROU I., KERSEY P. J., LOCKHART P., TAYLOR C., JAISWAL P. & WARE D. (2018). Gramene 2018 : Unifying comparative genomics and pathway resources for plant research. *Nucleic Acids Research*.

THE GENE ONTOLOGY CONSORTIUM (2014). Gene Ontology Consortium : going forward. *Nucleic acids research*, **43**(D1), D1049–1056.

VENKATESAN A., TAGNY NGOMPE G., HASSOUNI N. E., CHENTLI I., GUIGNON V., JONQUET C., RUIZ M. & LARMANDE P. (2018). Agronomic Linked Data (AgroLD) : A knowledge-based system to enable integrative biology in agronomy. *PLOS ONE*, **13**(11), 1–17.

Peuplement du jeu de données conférence

Thomas Cantié, Elodie Thiéblin, Cassia Trojahn

IRIT Institut de Recherche en Informatique de Toulouse, Toulouse, France
thomas.cantie.contact@gmail.com, elodie.thieblin@irit.fr, cassia.trojahn@irit.fr

Résumé : **Mots-clés** : Peuplement d'ontologie, Alignement d'ontologie, Évaluation d'alignements.

1 Introduction et travaux liés

Le jeu de données conférence proposé par Šváb Zamazal *et al.* (2005) est souvent employé pour de l'évaluation d'alignement d'ontologies, comme le montrent Zamazal & Svátek (2017). Ses alignements de référence utilisés dans l'OAEI (Ontology Alignment Evaluation Initiative), campagne annuelle d'évaluation de systèmes d'alignement, ont été déclinés sous plusieurs versions comme une version consensuelle (Cheatham & Hitzler, 2014) ou une version avec des alignements complexes (Thiéblin *et al.*, 2018a). Des approches d'alignement comme celles de Walshe *et al.* (2016) se basent sur des instances communes aux ontologies. Pour qu'elles puissent être évaluées et comparées aux autres approches d'alignement sur le jeu de données Conférence, il faut que celui-ci soit peuplé. D'autre part, dans la tâche OA4QA, Solimando *et al.* (2014) avaient succinctement peuplé quelques ontologies de Conférence pour évaluer des alignements sur de la réécriture de requêtes. Ce peuplement était limité à la portée des requêtes et était de fait difficilement réutilisable. Nous décrivons ici la méthodologie suivie pour peupler cinq ontologies du jeu de données Conférence et les jeux de données peuplés qui en résultent.

2 Méthodologie pour peupler le jeu de données

La méthodologie est basée sur des *questions de compétence pour alignement (CQAs)* qui sont définies comme des questions de compétence pouvant être couvertes par plusieurs ontologies (Thiéblin *et al.*, 2018b). Les CQAs définissent donc les besoins en connaissance devant être couverts (au mieux) par plusieurs ontologies et l'alignement entre elles. L'utilisation des CQAs dans le processus de peuplement assure qu'il soit homogène.

1. Création d'un ensemble de CQAs sur un scénario applicatif pour orienter l'interprétation des ontologies par les experts. Exemple de CQA : *Quels articles sont acceptés ?*
2. Création manuelle d'un format pivot (e.g., schéma JSON) pour couvrir les CQAs.
3. Pour chaque ontologie du jeu de données, créer des requêtes SPARQL INSERT pour traduire le format pivot. Chaque ontologie ne couvre pas forcément toutes les CQAs.
4. Instancier le format pivot avec des données réelles ou automatiquement générées.
5. Peupler les ontologies avec les instanciations du format pivot en utilisant les requêtes SPARQL INSERT de l'étape 3.
6. Appliquer un raisonneur aux ontologies peuplées. Si elles ne sont pas consistantes, changer son interprétation des ontologies et reprendre les étapes 3 à 5.

3 Jeu de données Conférence peuplé

La méthodologie a été suivie pour peupler cinq ontologies du jeu de données Conférence : *cmt, conference, confOf, edas, ekaw*. 152 CQAs ont été créées par un expert en suivant le

scénario réel d'organisation de Conférence sur l'édition de ESWC 2018 (Extended Semantic Web Conference) et étendues par exploration des ontologies. Le format pivot a été d'abord instancié avec les données du site Web d'ESWC 2018. Avec l'analyse de ces données, un script d'instanciation automatique du format pivot a été développé reprenant des statistiques telles que la proportion de membres du comité de programme auteur d'articles, etc. Le jeu de données, les instanciations du format pivot et le processus de peuplement sont disponibles [1].

En plus du peuplement avec les données d'ESWC, 6 jeux de données ont été générés pour proposer des ontologies partageant plus ou moins d'instances communes. Dans les jeux de données artificiels, chaque ontologie a été peuplée avec les données de 5 instanciations du format pivot (une instanciation du format pivot contient les informations pour l'organisation d'une conférence). Dans le jeu de données 0% toutes les ontologies ont été peuplées avec 5 instanciations du format pivot différentes. Dans le jeu de données 20%, les ontologies ont été peuplées avec 1 instanciation identique et 4 différentes. Les jeux de données 40%, 60%, 80% et 100% suivent la même logique. Le pourcentage qui sert de nom aux jeux de données est le pourcentage d'instanciations communes du format pivot utilisé dans le peuplement des ontologies. Comme la taille de chaque instanciation peut différer, le pourcentage d'instances communes entre deux ontologies varie. Par exemple, dans le jeu 20%, les instances *Articles scientifiques* communes aux ontologies représentent entre 7% des instances d'*Articles scientifiques* de l'ontologie *ekaw* et 11% des instances d'*Articles scientifiques* de l'ontologie *cmt*.

TABLE *1 – Nombre d'entités peuplées sur nombre d'entités total par ontologie. Nombre de CQAs couvertes par chaque ontologie.*

	cmt	conference	confOf	edas	ekaw
Classes	26 / 30	51 / 60	29 / 39	42 / 104	57 / 74
Obj. prop.	43 / 49	37 / 46	10 / 13	17 / 30	26 / 33
Data prop.	7 / 10	13 / 18	10 / 23	11 / 20	0 / 0
CQAs couvertes	46	90	67	60	84

Conclusion

Cinq ontologies du jeu de données Conférence ont été peuplées à l'aide de questions de compétence pour alignement. Sept jeux de données différents, ayant plus ou moins d'instances communes, résultent du peuplement. Ces jeux de données peuplés, couplés aux alignements de référence existants, peuvent servir à l'évaluation de systèmes d'alignements.

Références

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

SOLIMANDO A., JIMÉNEZ-RUIZ E. & PINKEL C. (2014). Evaluating ontology alignment systems in query answering tasks. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272*, p. 301–304.

THIÉBLIN E., CHEATHAM M., TROJAHN C., ZAMAZAL O. & ZHOU L. (2018a). The first version of the oaei complex alignment benchmark. In *ISWC Posters and Demos*.

THIÉBLIN E., HAEMMERLÉ O. & TROJAHN C. (2018b). Complex matching based on competency questions for alignment : a first sketch. In *OM 2018 - 13th ISWC workshop on ontology matching*.

WALSHE B., BRENNAN R. & O'SULLIVAN D. (2016). Bayes-recce : A bayesian model for detecting restriction class correspondences in linked open data knowledge bases. *International Journal on Semantic Web and Information Systems (IJSWIS)*, **12**(2), 25–52.

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

ŠVÁB ZAMAZAL O., SVÁTEK V., BERKA P., RAK D. & TOMÁŠEK P. (2005). Ontofarm : Towards an experimental collection of parallel ontologies. *Poster Track of ISWC*, **2005**.

1. `https://framagit.org/IRIT_UT2J/conference-dataset-population`

Cultivating Semantics for Data in Agriculture and Nutrition Recommendations from the RDA Agrisemantics Working Group

Sophie Aubin[1], Caterina Caracciolo[2] et Brandon Whitehead[3]

[1] INRA, UAR 1266 DIST Délégation Information Scientifique et Technique, Versailles, France
sophie.aubin@inra.fr
[2] Food and Agriculture Organization of the United Nations, Rome, Italy
caterina.caracciolo@fao.org
[3] CABI, Wallingford, UK
b.whitehead@cabi.org

Résumé : The poster reports on results of the activities carried within the Agrisemantics Working Group of the Research Data Alliance (RDA) (https://www.rd-alliance.org/groups/agrisemantics-wg.html). The group gathers together researchers and practitioners at the intersection between semantic technologies and agriculture. As a highly interdisciplinary domain, agriculture needs semantics as a possible response to data findability and interoperability. Yet, semantics remain unevenly adopted in this community partly because of unawareness, skill gaps, and lack of tooling usable by non-semantics experts to edit, map, and use semantic resources. The Agrisemantics WG has recently produced a set of recommendations to facilitate the use of semantics for Data on Agriculture and Nutrition, with the ultimate goal of improving data interoperability, discoverability and reusability. Based on an open dialog with a community of researchers and practitioners of semantic technologies, we have identified four basic points to address in order to make their use more widespread and effective. First, foster the development of a generic web-based framework to work with semantic resources. Such a framework should be adaptable according to: the task to perform, the domain of interest, the coverage within the domain, and the user competencies. Second, promote initiatives that support the reuse of existing resources, or their alignment. Third, promote the adoption of common metadata models for the description of semantic resources and alignments, and the use of global identifiers to facilitate reuse, usage tracking, and proper citation. Finally, promote courses on semantics and integrate them into curricula in all relevant disciplines– ie. computer sciences, data & information management and analytics, agronomy, bioinformatics, etc. These high level recommendations are further specified and translated into specific recommendation related to roles and activities in: data stewardship (or data management), software development, semantics professionals, and policy makers. Many of our recommendations either address or indirectly contribute to the implementation of FAIR principles. Our poster will show higher level recommendations together with the specific recommendations.

Mots-clés : Semantics, data interoperability, Research Data Alliance, good practices

Fouille de texte et fouille de graphe appliquées à la recherche d'experts

Stella Zevio, Guillaume Santini, Haïfa Zargayouna, Thierry Charnois

LIPN CNRS UMR 7030, Villetaneuse, France
zevio@lipn.univ-paris13.fr

La recherche d'experts est une problématique récurrente dans le milieu académique. En effet, les chercheurs doivent continuellement être assignés à des comités de programme, de recrutement, à des projets de recherche ou à des expertises de projet. Cette problématique est historiquement liée à celle du profilage d'experts (Lin *et al.*, 2017). Elle implique d'identifier des expertises (définies par Draganidis & Mentzas (2006) comme des compétences, connaissances, aptitudes ou comportements) et de les assigner aux individus adéquats (Bordea, 2013). Identifier les relations qu'un chercheur entretient au sein de la communauté scientifique est également essentiel pour définir son niveau d'expertise en prenant en compte une validation par les pairs. Notre objectif est donc d'identifier des ensembles d'experts (chercheurs) validés par leurs pairs et leur expertises (thématiques de publication) associées.

Afin d'automatiser cette tâche, les expertises et relations entre experts sont extraits automatiquement à partir de texte. Dans le milieu académique, les publications scientifiques recèlent de connaissances pertinentes pour construire les profils d'expert. Nous utilisons des méthodes classiques de fouille de texte (Khan *et al.*, 2017) pour extraire les thématiques de publication et les liens de collaboration scientifique entre chercheurs (citation, co-auteurs, *etc.*) à partir des publications scientifiques. Les connaissances extraites à partir du texte sont communément représentées sous forme de graphe attribué. L'originalité de notre approche consiste en l'adjonction d'une méthode de fouille de motifs exploitant les caractéristiques topologiques d'un graphe afin d'extraire de nouvelles connaissances.

La méthode de fouille de graphe que nous utilisons est l'abstraction de graphe. Elle est inspirée de l'analyse des réseaux sociaux et a déjà été utilisé pour la détection de k-communautés fréquentes (Soldano *et al.*, 2015). Partant du principe que les chercheurs et leurs expertises associées sont représentées sous la forme d'un graphe attribué, nous détectons des k-communautés fortement connectées dans le graphe. Les relations connues liant les auteurs ou les publications sont utilisées pour modéliser l'insertion d'un expert et de ses expertises dans une communauté scientifique et comme critère garantissant une certaine validation par les pairs. Ne seront considérées que les associations experts/expertises supportées par des structures connexes dans le réseau de relations scientifiques. Les structures connexes que nous cherchons à identifier s'appellent des cœurs.

L'identification des cœurs (*core*) d'un graphe est une approche classique d'exploration de la structure des graphes complexes. Le cœur d'un graphe est un sous-graphe maximal dont l'ensemble des sommets vérifient une propriété topologique. La première définition de cœur est celle du cœur *k-core* (Seidman, 1983) pour laquelle l'ensemble des sommets vérifient la propriété d'avoir un degré supérieur ou égal à k. Cette notion a été généralisée (Batagelj & Zaversnik, 2011) et permet la définition de nouveaux cœurs garantissant d'autres propriétés topologiques. Dans notre étude, nous nous intéressons à quatre propriétés topologiques.

Notre postulat est le suivant : si les experts sont fortement liés entre eux (par un réseau dense de relations de citation ou de co-publication par exemple), ils partageraient alors un ensemble d'expertises communes plus grand. En se focalisant sur les k-communautés fréquentes du graphe, notre hypothèse est que nous pourrions à la fois détecter des communautés d'experts, mais également les expertises communes maximales qu'ils partagent.

Une phase exploratoire est en cours sur un échantillon du corpus ACL Anthology (Bird *et al.*, 2008; Gábor *et al.*, 2016) constitué de 13322 publications scientifiques, publiées entre 1985 et 2008 sur les thématiques de la linguistique informatique et du traitement de la langue

naturel. De très nombreux résultats émergent, et la problématique consiste à identifier les paramètres et résultats les plus intéressants. La mise en place d'une méthode d'évaluation adaptée, automatique et rigoureuse nous paraît nécessaire à ce stade. Pour évaluer les résultats obtenus, nous ne disposons d'aucun *gold standard*. La dernière publication issue de notre échantillon du corpus ACL datant de 2008. Nous voulons donc en construire un en se basant sur le chapitre d'un livre de référence datant de 2010 et intitulé Information Extraction (Hobbs & Riloff, 2010) pour valider nos résultats. Pour faire cela, nous identifions les publications citées par le chapitre et référencées dans le jeu de données ACL. Pour chacune de ces publications, nous nous reportons à la partie du chapitre pour laquelle la publication est citée et nous étiquetons la publication avec les mots clefs trouvés dans le texte. Cet étiquetage des publications nous donne une référence à laquelle comparer les motifs énumérés par l'abstraction de graphe. Des 92 références que nous avons automatiquement extraites de ce chapitre, nous retrouvons automatiquement 37 publications issues de l'échantillon du corpus ACL analysé avec des traitements simples.

L'analyse préliminaire des résultats que nous avons obtenu montre que notre méthode d'extraction de concepts sémantiques à partir du texte pourrait également être améliorée. Certains des concepts sémantiques extraits ne sont pas très pertinent (par exemple *paper*, *method* ou *based*). À l'instar du système proposé par Osborne *et al.* (2013) les thématiques et relations extraites pourraient être liées à des concepts d'ontologies pour une meilleure interopérabilité sémantique et pour pouvoir généraliser les concepts identifiés. En perspective, nous souhaitons lier les concepts extraits automatiquement dans les résumés des publications à des concepts d'ontologie plus ou moins spécialisés, afin d'enrichir les motifs clos abstraits.

Références

BATAGELJ V. & ZAVERSNIK M. (2011). Fast algorithms for determining (generalized) core groups in social networks. *Adv. Data Analysis and Classification*, **5**(2), 129–145.

BIRD S., DALE R., DORR B. J., GIBSON B., JOSEPH M. T., KAN M.-Y., LEE D., POWLEY B., RADEV D. R. & TAN Y. F. (2008). The ACL Anthology Reference Corpus : a Reference Dataset for Bibliographic Research in Computational Linguistics.

BORDEA G. (2013). *Domain Adaptive Extraction of Topical Hierarchies for Expertise Mining*. PhD thesis.

DRAGANIDIS F. & MENTZAS G. (2006). Competency Based Management : a Review of Systems and Approaches. *Information Management & Computer Security*, **14**(1), 51–64.

GÁBOR K., ZARGAYOUNA H., BUSCALDI D., TELLIER I. & CHARNOIS T. (2016). Semantic Annotation of the ACL Anthology Corpus for the Automatic Analysis of Scientific Literature. In *LREC 2016*, Proceedings of the LREC 2016 Conference.

HOBBS J. R. & RILOFF E. (2010). Information Extraction. *Handbook of Natural Language Processing*, **2**.

KHAN S., LIU X., SHAKIL K. A. & ALAM M. (2017). A Survey on Scholarly Data : From Big Data Perspective. *Information Processing & Management*, **53**(4), 923–944.

LIN S., HONG W., WANG D. & LI T. (2017). A Survey on Expert Finding Techniques. *Journal of Intelligent Information Systems*, **49**(2), 255–279.

OSBORNE F., MOTTA E. & MULHOLLAND P. (2013). Exploring Scholarly Data With Rexplore. In *International Semantic Web Conference*, p. 460–477 : Springer.

SEIDMAN S. B. (1983). Network structure and minimum degree. *Social Networks*, **5**, 269–287.

SOLDANO H., SANTINI G. & BOUTHINON D. (2015). Local Knowledge Discovery in Attributed Graphs. In *International Conference on Tools with Artificial Intelligence (ICTAI)*, p. 250–257.

Rules-based decision support system and domain ontology for diabetes diagnosis

Dr. Dendani Nadjette[1] Allouani Rayene[2]

[1] Badji Mokhtar University,P.O.Box 12, Annaba, Algeria , LabGed Laboratory, Computer Science Department
[2] Badji Mokhtar University,P.O.Box 12, Annaba, Algeria , Computer Science Department
[1]n_dendani@yahoo.fr
[2]rayenneall@gmail.com

Abstract

Artificial intelligence must have access to objects, categories, properties and relations between all of them to implement knowledge engineering. Initiating common sense, reasoning and problem-solving power in machines is a difficult and tedious. AI decision support systems are used in multiple fields, such as the medical field. In which making a decision consists of posing a diagnosis and proposing a treatment. Various applications of decision-making support were developed in this domain. Those applications are intended to help medical workers (doctors, nurses...etc) in their process of decision making. It involves the use of powerful tools of Artificial Intelligence, such as the rules-based reasoning (RBR). Which stimulates the judgment and behaviour of a human or an organization that has expert knowledge and experience in a particular field by following a certain set of rules. Systems that use rules-based reasoning are also known as Expert Systems. An expert system is composed of a user interface, an inference engine and a knowledge base. The use of ontologies in the medical domain has seen a huge growth in recent years and was accompanied by a great success, which motivated us to create an ontology related to our field of work, the diagnosis of diabetes, to serve as our knowledge base. In this paper we proposed a rules-based decision support system for patients with diabetes. It uses the Rules-based technique for reasoning and a domain ontology for representing the knowledge to detect diabetes, its type, its seriousness and giving the appropriate care plan. This system helps both doctors and patients to check, analyse and repair solutions. It analyzes the symptoms of the patients and gives the exact types of diabetes, its seriousness level and the appropriate treatment for every patients. In addition to that, it offers an analysis of the data stored by the system, based on different factors such as: type of diabetes, age of the patient...

Keywords

Rules-based reasoning, Expert Systems, Ontology, Diabetes, Diagnosis, Decision-making, Artificial Intelligence, Medicine.

1 Introduction

Nowadays, diabetes is considered one of the most common diseases in most countries; In fact, according to the International Diabetes Federation, 425 million people were diagnosed with diabetes in the world in 2017, which represents about 5.5% of the world's population. Thus, it is expected that approximately 622 million people(8.1% of the world's population) would get diabetes comes 2040.

In the medical field, decision-making involves posing a diagnosis and determining the treatment, which is the stage where the medical worker determines the disease from which the patient suffers, using clinical symptoms, blood tests, radiological scans... etc. It is considered the core of any disease's treatment process.

The diagnosis is an intelligent act which is hardly programmable with classic techniques. Several studies have been conducted for the development of medical diagnosis methods based on Artificial Intelligence (AI) methods and techniques.There are several AI tools that lead to the realization of intelligent systems similar to the reasoning of experts in the field helping in the process of decision-making, among them the rules-based reasoning (Expert systems), which have proven their performance in the medical field, and since the use of ontologies in the medical field have seen a significant growth in the recent years, we were motivated to create an ontology related to our field of work, the diagnosis of diabetes.

The paper is organized as follows: Section 2 gives a theoretical background about Expert System, ontology, and the domain of application. The description of the system concept development is given in section 3. Section 4 presents description of the proposed architecture. Section 5 details the implementation of the system, while section 6 we discuss our work and give the system performance and conclude it in the sixth one.

2 Theoretical background:

The proposed approach combines AI paradigms previously cited, for instance Ontologies and Expert Systems.

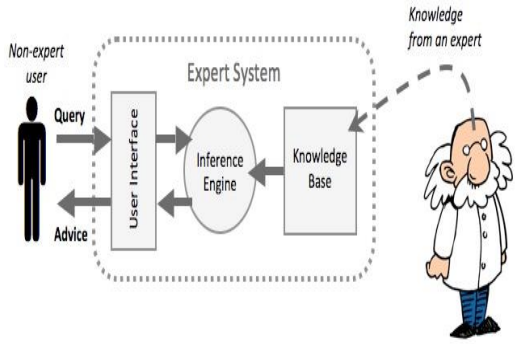
2.1 Ontology:

Gruber [1] defines an ontology as an explicit specification of a conceptualization. A conceptualization is an abstract view of the world that we hope to represent for a specific purpose. The representation of knowledge is based on a conceptualization including objects, concepts and relationships between them. Based on a technical perspective, the Sowa definition [2] considers an ontology as a specification of the kinds of entities that exist or may exist in certain domains or subjects. More formally, an ontology can be represented by a collection of nouns conceiving concepts and types of relations controlled in a partial order by the relation type / subtype. Several types of classifications are proposed

for ontologies based on the characteristic of the ontology's components. Few of the most important types of classifications are classification according to; Purpose, in which we find: application ontology and reference ontology, Expressiveness in which we find: heavyweight and lightweight ontologies, Specify in which we find: generic, core and domain ontologies [3].In this paper we have created domain ontology.

2.2 Expert System:

In general, an expert system is a tool capable of reproducing the cognitive mechanisms of an expert, in a particular field. More precisely, an expert system is a software capable of answering questions, making reasoning based on known facts and rules. It can be used as a tool for decision support[4]. The architecture of a typical expert system consists of several modules



that interact, which is detailed in **Fig.1**.

Fig 1: Architecture of an Expert System.

User Interface: serves to simplify the communication, it can use the question-answer form, the menu, the natural language etc.

Knowledge base: is the set of data that is used by the inference engine, it contains the knowledge about solving the problem.

Inference Engine: is a mechanism for inferring new knowledge from the system knowledge base. It is based on rules of inference that govern its operation. Its function is to answer a request from a user or a server to trigger a reflection defined by its inference rules that will use the knowledge base.

2.3 Diabetes:

Diabetes mellitus, often referred to as diabetes is a chronic disease that prevents the body from properly using sugar as a source of energy. It is caused by a lack or lack of use of a hormone called insulin which is produced by the pancreas. It allows glucose (sugar) to enter the cells of the body for use as a source of energy. There are four types of diabetes[5];

Type 1 diabetes: it affects about 6% of the diabetic population and often occurs during childhood, adolescence or early adulthood, rarely in older people. It is characterized by autoimmune destruction of beta cells, that produce insulin, causing total or partial insulin deficiency and effectively requiring the patient's daily administration of this hormone, people with type 1 diabetes therefore depends on daily injections of insulin or insulin pump to ensure their survival. Genetic predispositions are one of the main factors in the onset of the disease [6].

Type 2 diabetes: Previously called non-insulin-dependent diabetes or diabetes of maturity; Type 2 diabetes affects about 91% of the diabetic population. It appears at a later age, although it is progressing today towards an increasingly young population.. Unlike type 1 diabetes, type 2 diabetes is mostly asymptomatic [7].

Gestational diabetes: it is usually type 1 (insulin-dependent) and occurs in about 3-6% of pregnant women. It corresponds to transient glucose intolerance and appears in the third trimester of pregnancy but usually disappears after delivery. Gestational diabetes is the cause of a predisposition to type 2 diabetes [8].

Neonatal diabetes: it occurs in the first days or weeks of life. It is very rare, 250 to 500 000 newborns in Europe. As many boys as girls. It can be permanent (treatment can't be stopped) or transient (insulin can be stopped, usually before the age of 6 months, but recurrence of diabetes is possible at puberty or adulthood) [9].

3 System concept development:

3.1 Goals of proposed system:

Our research work proposed a rules-based decision support system for patients with diabetes. It is an artificial intelligence technique (Expert Systems) to detect diabetes and its type, its seriousness and giving the appropriate care plan. This system helps doctors and patients to check, analyze and repair solutions.

To solve an actual problem a set of rules is followed to reach a diagnosis then another set is followed to find an appropriate treatment[10]. Once both the diagnosis and the treatment are generated and validated, the case is stored in the ontology.

3.2 Some symptoms and rules followed to reach a diagnosis:

An effort has been made to find a co-relation between the initial cases and the input parameters/attributes of the cases to arrive at the relative importance of the parameters/ attributes. Weights have been assigned to each attribute on assumptions based on conclusions by experienced medical Professionals. **Table 1** shows the different weights of some of the symptoms against the various types of diabetes[11]:

Symptoms	*Weight*			
	D1	**D2**	**D3 (DG)**	**D4 (DN)**
Age (Patient old or young)	0.9	0.9	0.7	0.9
Gender	0.5	0.5	0.9	0.5
Polyuria-polydipsia syndrome	0.9	0.9	0.9	0.9
Weight loss	0.9	0.5	0.6	0.9
Polyphagy	0.9	0.8	0.7	0.5
Asthenia	0.9	0.8	0.8	0.9
Drowsiness	0.7	0.6	0.5	0.5
Blurry vision	0.7	0.6	0.5	0.5
Dehydration signes	0.6	0.7	0.5	0.8
Polypnoea	0.6	0.7	0.5	0.8
Diabetic ketoacidosis	0.8	0.6	0.5	0.8
Bad healing	0.5	0.7	0.5	0.6
Frequent infections	0.5	0.7	0.5	0.8
Obesity	0.5	0.9	0.7	0.6
Cardiac or vascular pathology	0.5	0.8	0.6	0.5
Family history in metabolic or obstetric syndromes	0.5	0.8	0.8	0.5
Personal or family history in autoimmune diseases.	0.8	0.5	0.6	0.9

TABLE 1: Symptoms and their weight values.

In order to reach a proper diagnosis the system follows rules, to name a few:

- **IF**((Age <=2y.o) **AND** (Random blood sugar level>=2g/l) **AND**(Asthenia=="yes")) **THEN** Diagnosis = Neonatal diabetes;
- **IF**((Pregnancy=="yes") **AND** (Fasting blood sugar level>=0.92g/l) **AND** (Polyuria-polydipsia=="yes")) **THEN** Diagnosis = Gestational diabetes;
- **IF**((HB1AC>=6.5%) **AND** (Age< 40y.o) **AND** (Antibodies=="yes") **AND** (Weight loss =="yes")) **THEN** Diagnosis= Type1 diabetes;
- **IF**((Postprandial blood sugar level>=2g/l) **AND** (Age>= 40y.o) **AND** (Polyphagia=="yes")) **THEN** Diagnosis= Type2 diabetes;

3.2 System architecture:

The proposed approach includes the use of ontologies to build models of general domain knowledge. The more knowledge is embedded into the system, the more effective it is expected to be. In this Expert system the Ontology play the role of a knowledge base in which the knowledge from the expert is stored. It plays an important role as a vocabulary to describe cases. In **Fig.2.**we show the proposed architecture which is composed of two functional components, domain ontology and the Expert system application.

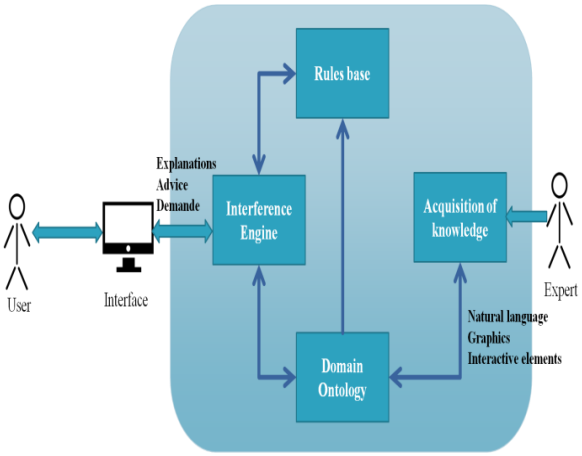


Fig.2. System Architecture.

Domain Ontology: The domain ontology is implemented to build general knowledge models and which includes vocabularies, concepts and relations for representing all knowledge concerning medical diagnosis of Diabetes, its type, its seriousness and its care plan[12].

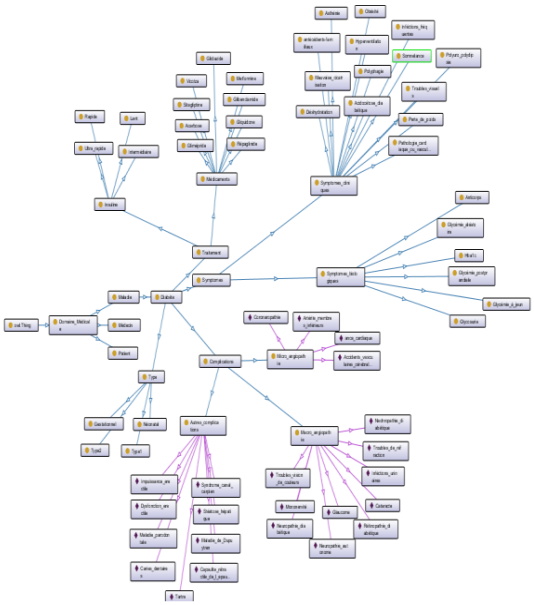


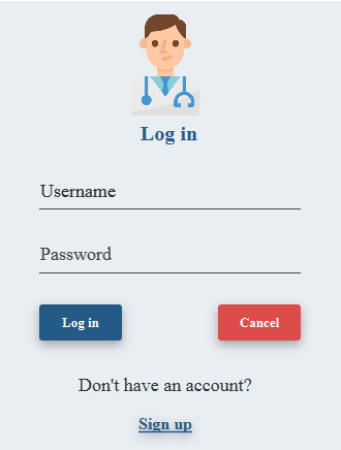
Fig.3. Graphical representation of the ontology's classes.

Expert System: This component is used for solving a diagnosis problem by following the rules that are in the IF ... THEN form.

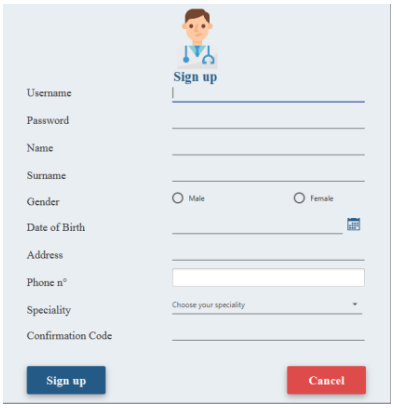
4 System implementation:

The realization of the system passed through different stages, carried out through a set of environments, tools and development libraries. To create the system we used the Java language in the environment: NetBeans IDE8.2, for the creation and modification of graphical interfaces, we used the Javafx library and the tool Gluon SceneBuilder 8.4.1. For the development of the ontology Protected tool 5.2.0 was used.

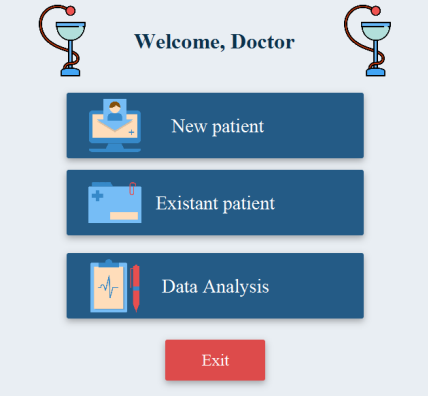
This screenshot represents the first interface of the system, it requires a username and password to log in, and if the user does not have an account, it allows them to sign up.



If the user clicks on the "Sign up" button, this window opens:

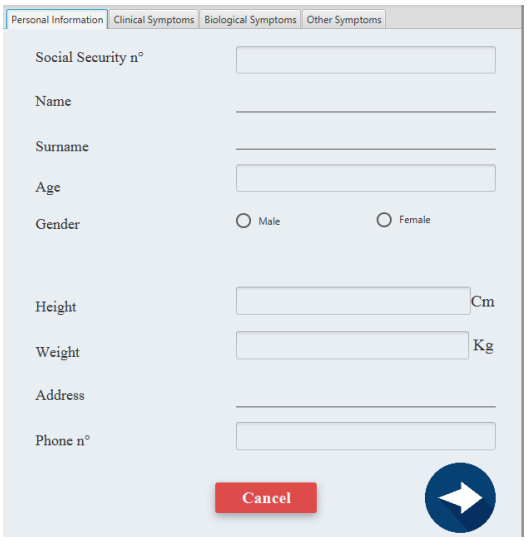


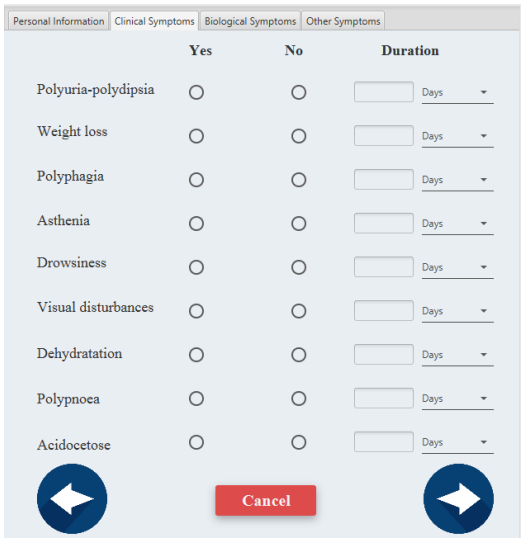
Once the user logs in, a main menu is displayed. This one allows him to add a new patient, to consult and possibly to modify an existing patient or to consult the analyzes of the data that the system contains.

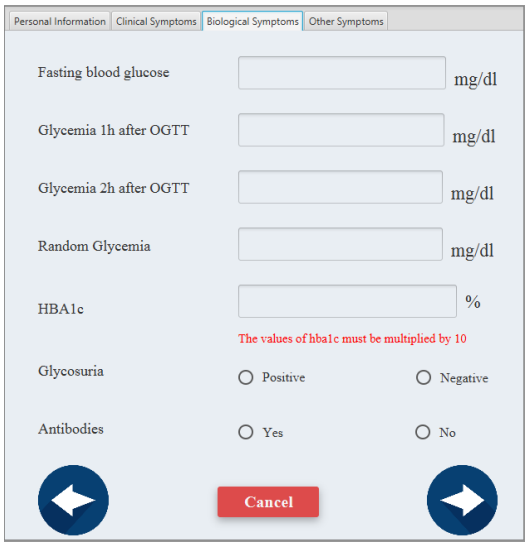


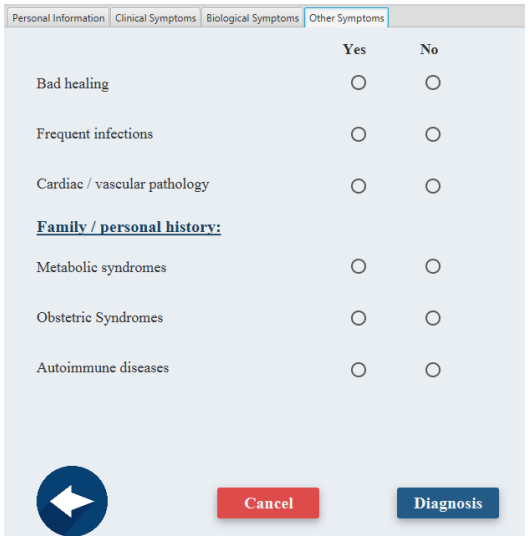
4.1 New Patient:

If the user chooses to add a new patient, this window - with multiple tabs- appears:

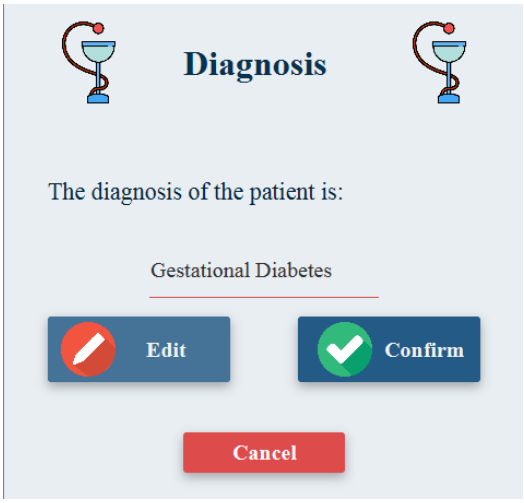




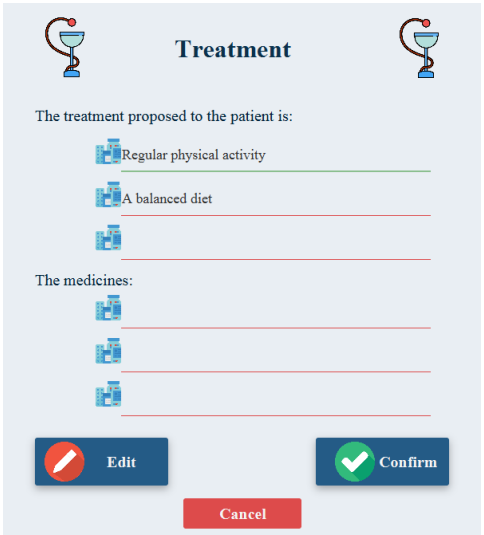




Once the user has completed all the necessary fields and presses the "Diagnosis" button, the system applies a function to find out whether the patient is diabetic or not and the type of diabetes with which he is suffering then the diagnostic window opens:



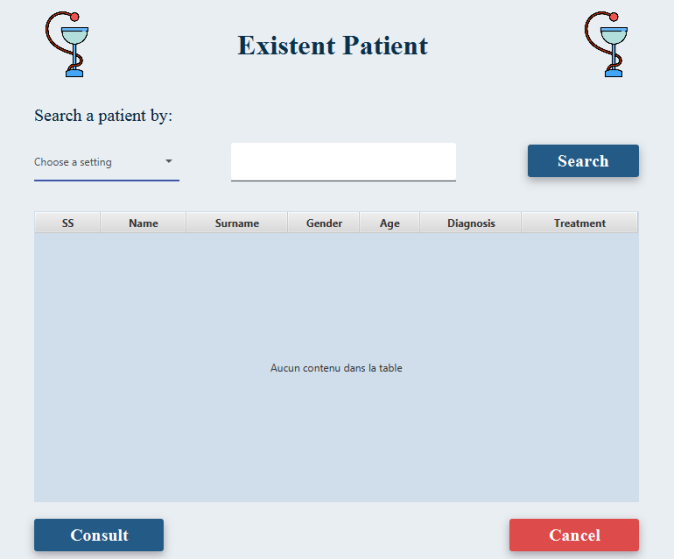
Then, the treatment window opens:



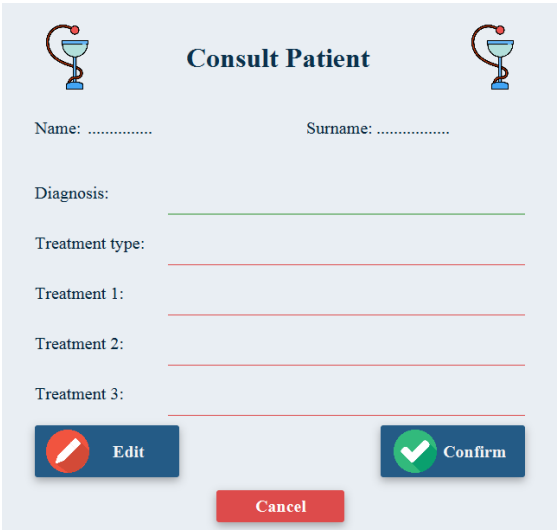
In case the treatment proposed by the system is not correct, the user can press the "Modify" button which allows him to modify the type of treatment and / or the proposed drugs, if there is no error the user validates the treatment. The system will then register the patient, his diagnosis and treatment in the database.

4.2 Existent Patient:

If the user chooses to consult existing patients this window that displays all the patients who have been viewed by this doctor opens:

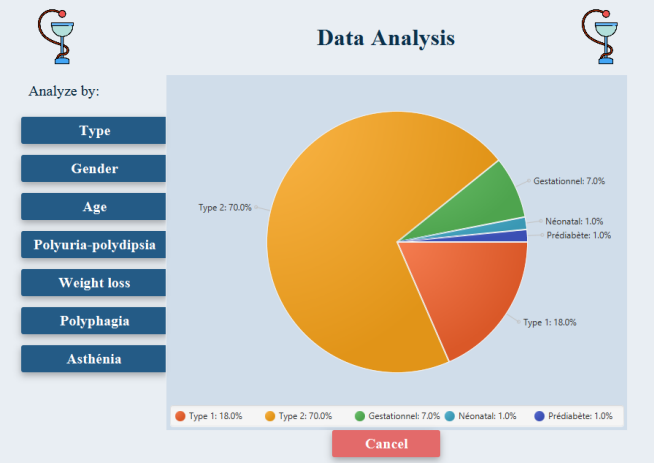


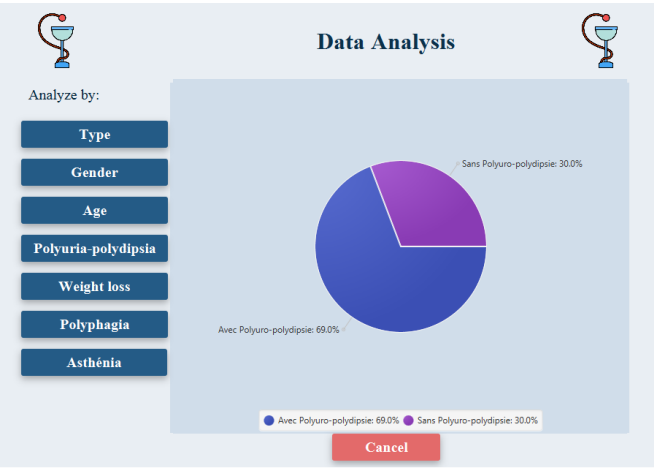
Once the user chooses a patient and clicks on the button "Consult" this window that allows him to edit the type of treatment, medication or even the diagnosis of the patient is opened:



4.2 Data Analysis:

If the user chooses to see the data analysis of the system, this window that allows him to analyze the data by type, age, sex, polyuro-polydipsia, weight loss, polyphagia or asthenia opens:





5 System Performance:

The system has been tested by a number of existing patients, obtained by a Data gathering at the Ibn Sina-Annaba UHC. Then, the results were compared to see the efficiency of this decision-support system. **Table 2** shows the obtained results;

	Cases introduced	Correct cases	Percentage
System results	**103**	**93**	**90.29%**
Data gathering results	**103**	**103**	**100%**

TABLE 2: System performance results.

6 Conclusion:

In this work, a tool has been realized that aims to help doctors to diagnose diabetes, its type and it seriousness and propose the appropriate treatment based on the clinical and biological symptoms of the patient introduced by the user (the doctor). Rules-based reasoning was used to obtain the diagnosis and the corresponding treatment.

To represent the knowledge needed for this diagnosis, a domain ontology was created, using the tool Protégé 5.2.0 [13][14].

The results provided by this tool have been compared and validated with the results obtained by experimentation at the Ibn Sina-Annaba UHC, and various other practices. It has been found that this system offers correct and efficient results at 90.29%.

The decision support system is a technique perfectly adapted to the medical reasoning and is very promising when she is applied in systems of help to the diagnosis[15].

References:

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

[2] J. F. Sowa. Ontologies for sharing knowledge. In manuscript of the invited speech to the terminology and knowledge Congress of Engineering (TKE '96), Vienna, 1996.

[3] Semantic web and ontology, Dhana Nandini, 2014.

[4] A. Barr and E.A. Frigenbaum, The Handboo of Artificial intelligence, Vol. 1, William Kaufmann, 1981.

[5] Standards of medical care in diabetes, American Diabetes Association—2018.

[6] http://www.diabetes.org/diabetescare

[7] http://www.nlm.nih.gov/medlineplus/diabetes.html

[8] http://www.medicinenet.com/diabetes_mellitus

[9] http://www.who.int/diabetes/en/index.html

[10] Diabetescare the journal of clinical and applied research and education
january 2018volume 41 | supplement 1 print issn 0149-5992 online issn 1935-5548 printed in the usa

[11] Dendani.N, Allouani.R "A decision support system for thediagnostis of the Diabetes disease" 11e plate-forme Intelligence Artifielle PFIA'2018 Journée IA & Santé Du 2 au 6 Juillet 2018

[12] Dendani. N .&all. (2017) "Case Base Reasoning(CBR) and Domain ontology to diagnose diabetes disease" 4thInternational conference on computational and experimental science and engineering (ICCESEN'2017)

[13] Protege,Ontology Editor and Knowledge Acquisition System. <http://protege.stanford.edu/>. (2009).

[14] H. Knublauch, R. W. Fergerson, N. F. Noy, and M. A. Musen. "The protege owl plugin: An open development environment for semantic web

applications". Semantic Web - Iswc 2004, Proceedings, 3298:229-243, (2004).

[15] Guessoum. S, Dendani.N .&all "OntoLung : a decision support system for the diagnosis of the Lung Cancer" 4thInternational conference on computational and experimental science and engineering (ICCESEN'2017)

Illumination3.0: A Semantic Annotation Platform Based on Ontology for Medieval Illuminations

Djibril Diarra[1] Rami Belkaroui[1] Martine Clouzot[2] Christophe Nicolle[1]

[1] CIAD Lab (Connaissances et Intelligence Artificielle Distribuées), University of Burgundy, Dijon, FRANCE
http://www.ciad-lab.fr/
[2] ArTeHis Lab (Archéologie Terre et Histoire), University of Burgundy, Dijon, FRANCE
http://artehis.u-bourgogne.fr/

dl4djibril@gmail.com, {rami.belkaroui, cnicolle}@u-bourgogne.fr, martine.clouzot@wanadoo.fr

Résumé

Les enluminures médiévales sont très riches en contenu sémantique illustré par des ornements, objets symboliques et animaux fantastiques, qui sont reliés par des relations sémantiques contextuelles et cela complique souvent leur compréhension. Dans ce papier, nous proposons un système d'annotation de ces images médiévales afin d'aider à leur compréhension par le plus grand nombre de personnes. Il est basé sur des modèles de représentation de connaissance (ontologie) et des techniques de l'intelligence artificielle.

Mots Clef

Ontologie, Enluminure médiévale, Système d'annotation, Raisonnement.

Abstract

Medieval illuminations are very rich in semantic content illustrated by ornaments, symbolic objects, fantastic animals and decorative initials, linked each to others with contextual semantic relations that make them difficult to understand. In this contribution, we propose an annotation's platform for those medieval images in order to help in highlighting their public understanding by the means of a knowledge representation model (ontology) and artificial intelligence techniques.

Keywords

Ontology, Medieval illumination, Annotation system, Knowledge inference.

1 Introduction

The development and generalization of the Internet increase the presence of cultural heritage collections on the Web through the digital online version of institutions such as the museum, archives, and libraries which manage them in real life. Those institutions should adapt by progressively integrating functions and technologies of the Web in order to move from a simple exhibition of digital representation of the collections to a semantic description of them through their semantic annotation.

An annotation consists to categorize the components within a corpus and identify relations between specific components [1, 9]. It can be applied on a textual corpus [1] in the identification of named entities [4], for example, or on images in order to link them and make easier their retrieval from repositories (the Web or bounded databases) as in [6, 11]. In this contribution, we present a semantic annotation's platform of an item of those cultural heritage collections: *medieval illuminations*.

Medieval illuminations (see an example in fig. 1) are images which were painted on parchment in the medieval manuscripts in the Middle Age [3, 2]. Those images contain semantic elements richly decorated with ornaments, illustrations, fantastic animals, and decorative initials making their layout complex and their semantic meaning difficult to understand. To overcome those understanding problems, a semantic annotation could be very helpful, but it raises another problem which is the full requirement of an expert for that annotation since these picture's contents are difficult to understand. In this paper, we propose an ontology based semantic annotation's system in order to help the experts and enable the not experts to annotate illuminations following the ontology's terms in opposition of most of the systems developed for cultural images' annotation such as the one described in [5] or the one in [10].

One can see in our platform a restriction of the annotators who should follow our proposed semantic ontology's lexicons in the annotation's process, but this limitation has the advantage of avoiding the use of other lexicons unrelated to our images. However, the ontology can be dynamically extended by adding new concepts and relations with the respect of the expert's advice. Moreover, our ontology contains inference's rules (logical axioms which enable to infer new statements from a set of declared statements within an ontology) which could help toward a semi-automatic objects recognition within an illumination through the use of

deep learning's algorithms.
In the rest of this paper, section 2 describes our built domain ontology (see section 2.2), the web application (see section 2.4) in implementation and an API[1](Application Programming Interface) which intermediates them (see section 2.3). And section 3 concludes and outlines our future works.

Figure 1: Medieval Illuminations illustrating donations scenes of luxurious illuminated manuscripts to the Burgundy Duke, Philippe the Good. *Brussels, Royal Library of Belgium, ms. 9243, folio 185 verso, Chronicles of Hainaut by Jean Wauquelin, 1446*

2 The System Illumination3.0

This section presents our semantic ontology based annotation's platform which contains three main components: the ontology, API, and Web application. We describe them in an architectural organization in the subsections below.

2.1 Architecture of the System

The three components of *Illumination3.0* are structured in an architectural view as shown in fig. 2. It illustrates four steps for our platform's realization. The first (black dashed rectangle, *step 1*) indicates the ontology's creation. The second (yellow dashed rectangle, *step 2*) and third (red dashed rectangle, *step 3*) indicate the communications of the API with the ontology by a side and with the Web application by the other. And the fourth (green dashed rectangle, *step 4*) indicates the effective annotation's process of illuminations within the Web application.

2.2 The Ontology

The first component of *Illumination3.0* is a formal ontology (see fig. 3) which semantically describes the illuminations' contents. An ontology is a formal and explicit knowledge representation model [7]. We named the one

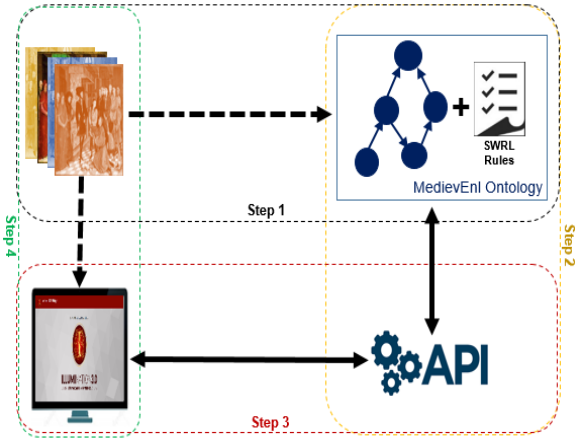


Figure 2: The architecture of illumination3.0

we create for our medieval illuminations (those of the Burgundy duke) by *MedievEnl*. We create it by analyzing some of those illuminations, extracting their contents by the help of an expert, and organizing them with ontological terms (classes, relations, individuals and axioms). Beyond a simple terminological organization, our ontology contains inference's rules which enable to reason about its contents. These rules are expressed in SWRL[2] language. They improve the expressivity power of the built ontology. The space of this paper is bounded for our ontology's full detail (for more specification about its components, see `http://medievenl.ontology.checksem.fr`). An ongoing article is submitted about it.

2.3 The Application Programming Interface (API)

The second component is an API (see an interface in fig. 4) which enables to manage the ontology. It is a combination an OWL API[3] and a Pellet reasoner[4]. The former enables to explore the ontology's components and extend the ontology by adding new concepts or individuals. And the latter enables to reason about the ontology, through its contained inference's rules, in order to infer new knowledge from the explicitly described ones. The choice of Pellet reasoner is because of its simplicity and free availability, but all other reasoner would be suitable.

2.4 The web application for annotation

The third component is a Web application whose an interface is shown in fig. 5. It is linked to the API in order to get knowledge from and put them into the ontology. Con-

[1] API is the acronym for Application Programming Interface, which is an intermediate software that allows two applications (or modules of an application) to talk each to other.

[2] SWRL, for Semantic Web Rules Language, is an inference's rules construction's language based on the logic of predicates (see [8])

[3] The OWL API is a Java API and reference implementation for creating, manipulating and serializing OWL ontologies. OWL (Ontology Web Language) is an ontology development language. It is a recommendation of 3WC (World Wide Web Consortium)

[4] Pellet reasoner is an open source Java based OWL 2 reasoner. It can be used in conjunction with both other reasoners and OWL API libraries.

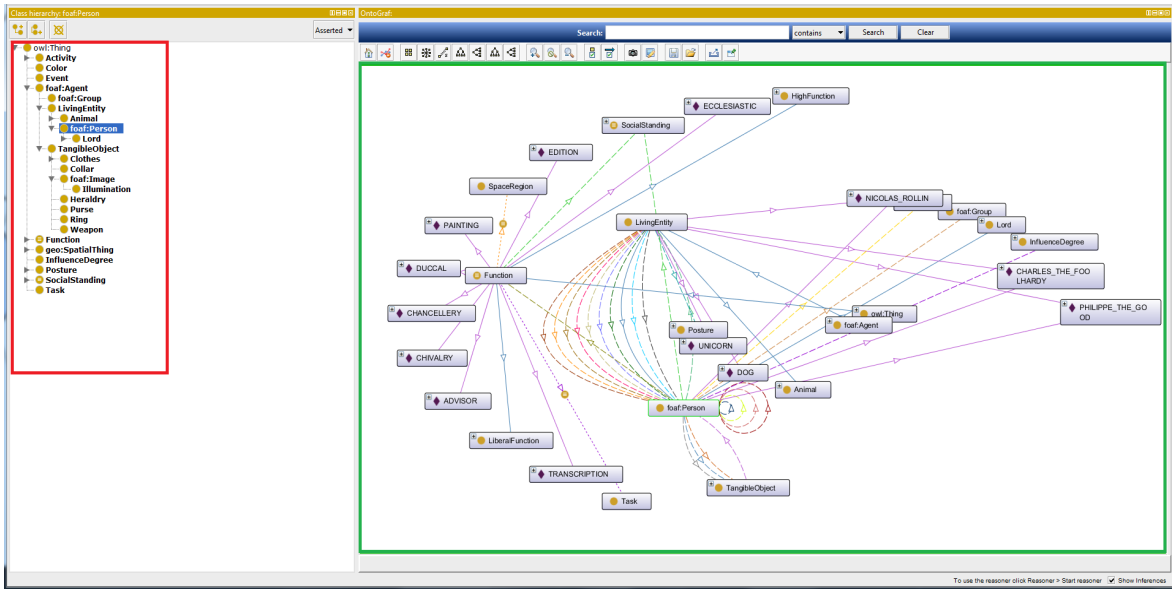


Figure 3: A view of our ontology. The red frame, in the left, shows a view of its concepts and the green one, in the right, shows a graphical visualization of some of them.

Figure 4: The API (Application Programming Interface) of *Illumination3.0*

cretely, it allows to upload an illumination, frame its important elements, annotate each of them as an individual of a concept in the ontology and connect those individuals with relations following the ones which exist in the ontology. For example, if a concept A is linked to B with a relation r in the ontology and a user annotates in an uploaded illumination an element a_1 as an individual of A and b_1 as an individual of B then s/he can link a_1 and b_1 par r.

If a concept or relation does not exist in the ontology then the user can add it to the terminological box (TBox) of the ontology, but these features are restricted to experts in order to add only consistent concepts and relations. Therefore, the annotation's process enables to populate the ontology by filling its assertional box (ABox) and extend it by adding new classes and relations to its Tbox.

Fig. 5 shows an interface in which one can see an illumination uploaded in the left half. Some entities in the image are framed with contoured rectangle. In the right half, some relations are indicated between those entities. We can see for example :

- **Individual annotated**: `Philippe le Bon` who is individual of concept `Person` in the ontology, `Chaise` which is individual of the concept `TangibleObject`, etc.
- **related individuals**: `Philippe le Bon|assis sur|Chaise` to state the relation `assoir` between the individuals `Philippe le Bon` and `Chaise`; `Wauquelin|offre|Manuscript` to state the relation `offrir` between `Wauquelin` and `Manuscript`; etc.

3 Conclusion and Perspectives

In this contribution, we describe an ontology based annotation's platform for medieval illuminations. It is a computerized system of an ongoing thesis project. We present the general architecture of the platform and detail its components which are a semantic ontology for the illuminations, an API for this ontology's easy management by a web application which completes the architecture. This latter component enables to annotated images following the ontology's terms and the API's reasoner allows to infer new knowledge from annotated entities.

As future works, we are implementing some features which take into account that reasoning process in the web appli-

Figure 5: An annotated illumination in our platform (`http://illumination2.checksem.fr/illumination/app/accueil`)

cation. Moreover, we are working on a deep learning algorithm in order to help in the recognition of an illumination content based on our ontology.

Acknowledgements

This work is funded by the Embassy of France in Bamako and the government of the Republic of Mali for the co-funding of a thesis in the University of Burgundy Franche-Comte (UBFC). We would like to thank all these institutions. We would also like to thank Lise Kabbache for help in the implementation.

References

[1] Benzitoun, C., Dister, A., Gerdes, K., Kahane, S., Marlet, R.: annoter du des textes tu te demandes si c'est syntaxique tu vois. In: 28th International Conference on Lexis and Grammar (LGC 2009). vol. 4, pp. 16–27. Presses de l'Université de Bergen (2009)

[2] Calkins, R.G.: Illuminated books of the middle ages. Thames and Hudson London (1983)

[3] De Hamel, C.: A history of illuminated manuscripts. Phaidon Press London (1994)

[4] Dutrey, C., Clavel, C., Rosset, S., Vasilescu, I., Adda-Decker, M.: Quel est l'apport de la détection d'entités nommées pour l'extraction d'information en domaine restreint ? (what is the contribution of named entities detection for information extraction in restricted domain?) [in french]. In: Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN. pp. 359–366 (2012)

[5] Foys, M., Bradshaw, S.: Developing digital mappaemundi: an agile mode for annotating medieval maps. Digital Medievalist **7** (2012)

[6] Ghosh, M.S., Bandyopadhyay, S.K.: A proposed method for semantic annotation on social media images. International Journal Of Engineering And Computer Science **6**(6) (2017)

[7] Gruber, T.: What is an ontology. WWW Site http://www-ksl. stanford. edu/kst/whatis-an-ontology. html (accessed on 07-09-2004) (1993)

[8] Horrocks, I., Patel-Schneider, P., Boley, H., Tabet, S., Grosof, B., Dean, M.: Swrl: A semantic web rule language combining owl and ruleml. W3C Member submission **21**, 1âĂŞ20 (01 2007)

[9] Kiryakov, A., Popov, B., Terziev, I., Manov, D., Ognyanoff, D.: Semantic annotation, indexing, and retrieval. Web Semantics: Science, Services and Agents on the World Wide Web **2**(1), 49–79 (2004)

[10] Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: a database and web-based tool for image annotation. International journal of computer vision **77**(1-3), 157–173 (2008)

[11] Sarwar, S., Qayyum, Z.U., Majeed, S.: Ontology based image retrieval framework using qualitative semantic image descriptions. Procedia Computer Science **22**, 285–294 (2013)

PADI-web: un système automatique multilingue pour la veille sanitaire internationale en santé animale

Sarah Valentin[1,2], Julien Rabatel[3], Elena Arsevska[1,3], Sylvain Falala[1,3], Jocelyn de Goer[4], Alizé Mercier[1,3], Renaud Lancelot[1,3], Mathieu Roche[2,3]

[1] UMR ASTRE, Univ. Montpellier, Cirad, INRA, Montpellier, France
sarah.valentin@cirad.fr
[2] UMR TETIS, Univ. of Montpellier, AgroParisTech, Cirad, CNRS, Irstea, Montpellier, France
[3] Cirad, Montpellier, France
[4] INRA, UMR EPIA, Clermont-Ferrand, France

Mots-clés : Santé Animale, Intelligence Epidémiologique, Web, Text Mining.

Conférence visée : Ingénierie des Connaissances (IC)

Détails techniques pour la démonstration : Accès à Internet

1 Introduction

La veille en santé animale a pour objectif l'alerte précoce vis-à-vis de dangers sanitaires connus ou émergents. Elle repose sur le recueil, le suivi et l'analyse quotidienne d'informations issues de sources officielles, telles que l'Organisation mondiale de la santé animale (OIE), et de sources non-officielles telles que les médias ou les réseaux sociaux (Hartley *et al.* (2010)). Plusieurs systèmes de biosurveillance, tels que MedISys (Mantero *et al.* (2011)), GPHIN (Blench (2008)) ou HealthMap (Freifeld *et al.* (2008)), sont ainsi dédiés à l'acquisition et à la diffusion de données issues de sources informelles. Ces systèmes s'intéressent à un large éventail de risques sanitaires (maladies infectieuses humaines, animales ou végétales, risques environnementaux, etc.), mais aucun d'entre eux n'est spécifiquement dédié à la santé animale. De plus, tous reposent sur une modération humaine à une ou plusieurs étapes de leur processus. Dans ce contexte, nous présentons PADI-web [1] (Platform for Automated extraction of Disease Information from the web), un outil de biosurveillance des médias digitaux pour la détection de foyers de maladies animales (Arsevska *et al.* (2018)). PADI-web est intégré dans la thématique de Veille sanitaire internationale, au sein de la plateforme d'Epidémiosurveillance en santé animale [2] (plateforme ESA). Depuis sa première version, dédiée à la veille de sources en anglais, PADI-web a été enrichi d'un nouveau classifieur reposant sur de l'apprentissage automatique et intègre les documents multilingues.

2 PADI-web : de la collecte d'articles à l'extraction d'information

PADI-web repose sur 4 étapes successives permettant d'extraire des informations épidémiologiques à partir du contenu d'articles relatifs à des événements infectieux en santé animale.

1. https ://padi-web.cirad.fr/en/
2. https ://www.plateforme-esa.fr/

IC 2018

2.1 Collecte des articles

L'aspiration des articles est effectuée quotidiennement et de manière automatique via l'agrégateur Google News, grâce à des requêtes intégrées sous la forme de flux RSS. Ces flux sont des combinaisons booléennes de mots-clés développées par une approche combinant l'extraction automatique de termes et la sollicitation d'avis d'experts (Arsevska *et al.* (2016)). Deux types de requêtes sont actuellement implémentés dans PADI-web. Les requêtes spécifiques incluent le nom d'une maladie (par exemple, « avian flu OR avian influenza OR bird flu »), et visent à détecter les événements vis-à-vis de maladies d'intérêt. Les requêtes non-spécifiques consistent en une combinaison de signes cliniques et de noms d'hôtes (par exemple, « abortions AND cows »), et permettent de détecter des événements non-prédéfinis.

2.2 Nettoyage du contenu et traduction

Le contenu des articles aspirés est nettoyé afin d'en supprimer les éléments inutiles (images, hyperliens, publicité, etc.), puis est enregistré dans une base de données accompagné des métadonnées de l'article (nom de la source, date de publication et titre). PADI-web filtre les éventuels doublons en comparant l'url de chaque nouvel article à ceux déjà existants dans la base de données. Les étapes de classification et d'extraction d'information reposant sur des modèles appris en anglais, tous les articles aspirés en une autre langue que l'anglais sont préalablement traduits. La langue source est détectée grâce à la librairie *langdetect* (Python) et la traduction repose sur l'API Translator du système Microsoft Azure. Sur une période de 3 mois, l'intégration des requêtes multilingues a permis d'augmenter le nombre d'articles pertinents de 131% pour la peste porcine africaine (207 articles en anglais, 272 traduits), de 47% pour l'influenza aviaire (212 en anglais, 99 traduits) et de 67% pour la fièvre aphteuse (104 en anglais, 174 traduits).

2.3 Classification

L'étape de classification est une étape cruciale dans le processus de PADI-web, car elle permet de filtrer la quantité d'articles qui seront présentés à l'utilisateur en rejetant les articles non-pertinents (non liés à un danger sanitaire). Le classifieur de PADI-web est issu d'un apprentissage automatique supervisé. Une sélection de différents modèles est entrainée une fois par jour sur un corpus d'apprentissage. Le modèle qui obtient les meilleures performances est sélectionné pour la classification des nouveaux articles (actuellement, il s'agit d'un classifieur de type Random Forest, qui obtient une exactitude (*accuracy*) moyenne de 0.97 en validation croisée). Le corpus d'apprentissage est un corpus annoté de 600 articles (200 articles pertinents et 400 articles non pertinents), pouvant être directement enrichi par l'utilisateur. A partir de l'interface, l'utilisateur peut en effet attribuer une classe à chaque nouvel article, indépendamment de la classe attribuée par le classifieur. Cette fonctionnalité permet de corriger les éventuelles erreurs de classification et d'augmenter facilement le jeu d'apprentissage. De plus, le module est générique : l'utilisateur peut créer autant de nouvelles tâches de classification que nécessaire (sous condition d'inclure un jeu de données annotées pour l'apprentissage). Les classes correspondant à chaque tâche de classification sont attribuées indépendamment les unes des autres par le classifieur.
Depuis sa mise en fonctionnement en février 2016, PADI-web a aspiré plus de 66 000 articles [3], dont 15 000 articles classés comme pertinents. Un échantillon de 100 articles aléatoirement sélectionnés dans la base de donnée de PADI-web a été manuellement évalué par deux épidémiologistes. L'exactitude (*accuracy*) sur cet échantillon est de 0.92.

2.4 Extraction d'information

La dernière étape de PADI-web consiste en l'extraction des indicateurs épidémiologiques dans le contenu des articles pertinents. Ce module est issu d'un apprentissage supervisé dé-

3. Les requêtes multilingues ayant été intégrées récemment, elles ne sont pas comptabilisées.

taillé et évalué par Arsevska *et al.* (2018). Brièvement, les noms de maladie, les hôtes et les symptômes sont détectés grâce à un dictionnaire créé manuellement et régulièrement enrichi, prenant en compte les synonymes pour chaque type d'hôte ou de maladie. Les localisations et les dates sont extraites respectivement grâce au gazetier GeoNames (Ahlers (2013)) et à HeidelTime, un système d'étiquetage d'expressions temporelles à base de règles (Strotgen & Gertz (2010)). Le principe mis en oeuvre est détaillé par Arsevska *et al.* (2018).

3 Interface de PADI-web

3.1 Recherche d'information

Les articles stockés dans la base de données de PADI-web sont consultables via une interface dédiée. Par défaut, les 10 derniers articles aspirés et classés comme pertinents sont affichés. Un large choix de filtres permet à l'utilisateur d'effectuer des recherches plus détaillées. Les articles peuvent être filtrés en fonction de différents attributs tels que la date de leur publication, leur classe (pertinent ou non pertinent) ou encore le nom de leur source. L'utilisateur peut également effectuer sa recherche sur la base du contenu des articles en utilisant les entités épidémiologique extraites (maladie, hôte, etc.) ou en recherchant un mot ou expression de son choix dans le titre ou le corps de l'article.

3.2 Visualisation et annotation

L'utilisateur peut accéder aux métadonnées, aux informations extraites et au contenu de chaque article des résultats d'une requête (Figure 1). Les entités extraites sont listées dans un encart et identifiées dans le texte avec une icône spécifique de chaque type afin de faciliter la visualisation des informations essentielles. Pour chaque entité, une fenêtre contenant des informations complémentaires peut être affichée. Un lien vers Google Maps est associé à chaque entité géographique. A partir de cette interface, l'utilisateur peut manuellement annoter la pertinence de l'article et des entités extraites. Les annotations sont automatiquement enregistrées et prises en compte lors des requêtes ultérieures.

3.3 Exports

Les résultats issus des requêtes peuvent être exportés sous différents formats. Le nombre d'articles correspondant à la requête en fonction du temps peut être visualisé par un histogramme, en utilisant plusieurs niveaux d'agrégation temporelle (par jour, mois ou année). L'utilisateur peut également exporter le jeu de données contenant les entités épidémiologiques extraites, en choisissant parmi différents formats (csv, json ou xls).

4 Conclusion

Nous proposons un outil de biosurveillance dédié à la veille en santé animale et adapté à une utilisation quotidienne par les épidémiologistes. Outre sa spécificité vis-à-vis du domaine vétérinaire, PADI-web repose sur des approches issues d'apprentissage automatique et de fouille de texte permettant de produire des données structurées et directement exploitables par les experts. L'interface permet à l'utilisateur de personnaliser ses requêtes et d'accèder rapidement aux informations pertinentes. Nous envisageons d'enrichir PADI-web d'un module d'extraction de signaux faibles afin d'identifier des informations épidémiologiques fines, telles que les mesures de lutte et de prévention ou les états d'alerte.

IC 2018

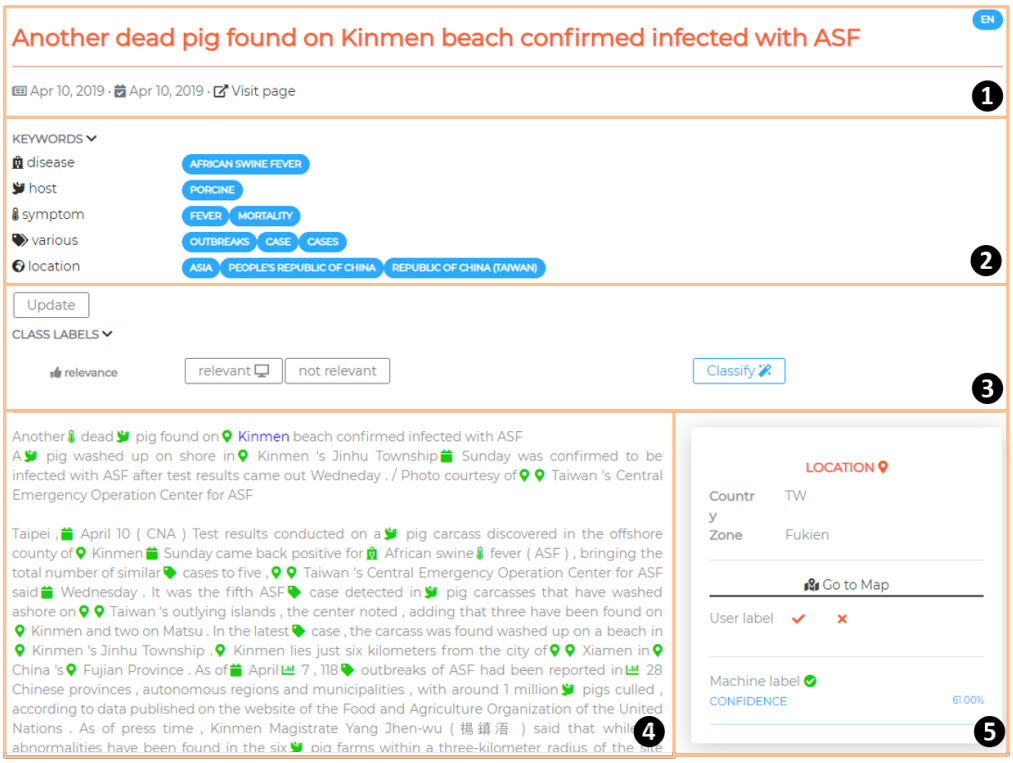


FIGURE *1 – Visualisation d'un article traité par PADI-web, contenant 1. les métadonnées de l'article (titre, date de publication, lien url vers l'article source), 2. la liste des mots-clés tagués, 3. la classe prédite par le classifieur, 4. le texte nettoyé avec les entités épidémiologiques extraites et 5. les informations liées à l'entité géographique sélectionnée 'Kinmen'.*

Références

AHLERS D. (2013). Assessment of the Accuracy of GeoNames Gazetteer Data. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*, p. 74–81, New York, NY, USA : ACM.

ARSEVSKA E., ROCHE M., HENDRIKX P., CHAVERNAC D., FALALA S., LANCELOT R. & DUFOUR B. (2016). Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web. *Computers and Electronics in Agriculture*, **123**, 104–115.

ARSEVSKA E., VALENTIN S., RABATEL J., DE GOËR DE HERVÉ J., FALALA S., LANCELOT R. & ROCHE M. (2018). Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System. *PLOS ONE*, **13**(8), e0199960.

BLENCH M. (2008). Global public health intelligence network (GPHIN). In *8th Conference of the Association for Machine Translation in the Americas*, p. 8–12.

FREIFELD C. C., MANDL K. D., REIS B. Y. & BROWNSTEIN J. S. (2008). HealthMap : Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. *Journal of the American Medical Informatics Association*, **15**(2), 150–157.

HARTLEY D., NELSON N., WALTERS R., ARTHUR R., YANGARBER R., MADOFF L., LINGE J., MAWUDEKU A., COLLIER N., BROWNSTEIN J., THINUS G. & LIGHTFOOT N. (2010). The landscape of international event-based biosurveillance. *Emerging Health Threats Journal*, **3**(0).

MANTERO J., BELYAEVA J., LINGE J., EUROPEAN COMMISSION, JOINT RESEARCH CENTRE & INSTITUTE FOR THE PROTECTION AND THE SECURITY OF THE CITIZEN (2011). *How to maximise event-based surveillance web-systems : the example of ECDC/JRC collaboration to improve the performance of MedISys.* Luxembourg : Publications Office. OCLC : 870614547.

STROTGEN J. & GERTZ M. (2010). HeidelTime : High Quality Rule-Based Extraction and Normalization of Temporal Expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, p. 321–324.