AfIA

Association française
pour l'Intelligence Artificielle

IC

Journées francophones d'Ingénierie des Connaissances

PFIA 2022

Table des matières

Fatiha Saïs
Éditorial . . . 6

Comité de programme . . . 7

Session 1 : Extraction d'informations et graphes de connaissances (1) . . . 9

Viet-Phi Huynh, Jixiong Liu, Yoan Chabot, Frédéric Deuzé, Thomas Labbé, Pierre Monnin, Raphaël Troncy
DAGOBAH : Annotation sémantique de données tabulaires par comparaison du contexte des tables et d'un graphe de connaissances . . . 10

Lucie Cadorel, Andrea Tettamanzi et Alicia Blanchi
Geospatial Knowledge in Housing Advertisements : Capturing and Extracting Spatial Information from Text . . . 20

Session 2 : Raffinement de graphes de connaissances - enrichissement et validation . . . 22

Franck Michel, Florence Amardeilh, Robert Bossy, Catherine Faron et Catherine Roussey
Alignement entre sources : cas d'usage des plantes cultivées . . . 23

Shadi Baghernezhad-Tabasi, Loïc Druette, Fabrice Jouanot, Céline Meurger et Marie-Christine Rousset
IOPE : Interactive Ontology Population and Enrichment Guided by Ontological Constraints . 33

Elodie Thiéblin, Ollivier Haemmerlé et Cássia Trojahn
Évaluation automatique d'alignements complexes : une approche basée sur des instances . . . 35

Thomas de Groot, Joe Raad et Stefan Schlobach
Analysing Large Inconsistent Knowledge Graphs using Anti-Patterns . . . 45

Session 3 : Résultats d'études et d'états de l'art en ingénierie des connaissances . . . 47

Gilles Kassel
Plaidoyer pour des ontologies épistémiques . . . 48

Beatrice Markhoff et Arnaud Soulet
Où sont les termes ? . . . 56

Yousouf Taghzouti, Antoine Zimmermann et Maxime Lefrançois
Négociation de contenu sur le Web : un état de l'art . . . 64

Session 4 : Modélisation de connaissances complexes (1) . . . 71

Maxime Lefrançois, Raul Garcia Castro, Maria Poveda-Villalon et Omar Qawasmeh
Apports des méthodologies et techniques de développement logiciel pour l'ingénierie des ontologies : Retour d'expérience des contributions au développement de l'ontologie ETSI SAREF . . . 72

Nathalie Aussenac-Gilles, Catherine Comparot, Antoine Dupuy, Nabil El Malki, Ronan Tournier, Ba-Huy Tran et Cassia Trojahn
eCISE-OWL : Représentation OWL du Schéma de Données de l'Environnement Commun d'Échange d'Information pour le Domaine Maritime . . . 82

Ba-Huy Tran, Nathalie Aussenac-Gilles, Catherine Comparot et Cassia Trojahn
Intégration sémantique de données Raster pour l'observation de la Terre sur des unités territoriales . . . 92

Session 5 : Explicabilité et interprétabilité dans les graphes de connaissances . . . 94

Nicholas Halliwell, Fabien Gandon et Freddy Lecue
Evaluation d'explications pour la prédiction de liens dans les graphes de connaissances par des réseaux convolutifs . . . 95

Ismail Harrando et Raphaël Troncy
Utiliser les connaissances du sens commun pour la découverte des topics intérpretables 97

Antonia Ettorre, Anna Bobasheva, Franck Michel et Catherine Faron
Stunning Doodle : un outil pour la visualisation et l'analyse conjointe de graphes de connaissances et leurs plongements 99

Session 6 : Apprentissage automatique, ontologies et graphes de connaissances 101

Lucas Simonne, Nathalie Pernelle, Fatiha Saïs et Rallou Thomopoulos
Découverte de règles causales dans les graphes de connaissances à l'aide de plongements dans les graphes 102

Cédric Baudrit, Franck Michaud et Christophe Fernandez
Identifier et reconnaitre des essences de bois à l'aide d'un réseau Bayésien basé sur des indicateurs macroscopiques 112

Melanie Munch, Patrice Buche, Cristina Manfredotti, Pierre-Henri Wuillemin et Hélène Angellier-Coussy
Une approche d'ingénierie inverse combinant ontologies et modèles relationnels probabilistes : application aux emballages bio-composites 120

Session 7 : Graphes de connaissances et temporalité 130

Lucas Bourel, Nathalie Hernandez, Nathalie Aussenac-Gilles et William Charles
HHT : une ontologie modulaire pour représenter l'évolution des territoires en Histoire 131

Nassira Achich, Fatma Ghorbel, Fayçal Hamdi, Elisabeth Métais et Faiez Gargouri
Traitement des données temporelles certaines et incertaines en OWL 2 : Approche basée sur la théorie des probabilités 137

Jacques Hilbey, Xavier Aimé and Jean Charlet
Représentation des connaissances médicales temporelles au moyen d'ontologies 147

Session 8 : Modélisation de connaissances complexes (2) 153

Helen Mair Rawsthorne, Nathalie Abadie, Eric Kergosien, Cécile Duchêne et Eric Saux
ATLANTIS : Une ontologie pour représenter les Instructions nautiques 154

Olivier Inizan, Vincent Fromion, Anne Goelzer, Fatiha Saïs et Danai Symeonidou
Une ontologie pour organiser des données biologiques : la contribution des modèles mathématiques 164

Session 9 : Ontologies et raisonnement pour les systèmes complexes 169

Fabien Amarger, Nicolas Chauvat et Elodie Thiéblin
OWL2YAMS : créer une application CubicWeb à partir d'une ontologie OWL 170

Olivier Poitou et Claire Saurel
Éviter l'échec des systèmes complexes : en construire collectivement une représentation formelle utile 176

Kevin Cousot, Thibaud Sanchez, Antoine Nguyen, Anthony Calpas, Ghislaine Martinez et Cédric Lopez
ElvirIA-P : génération d'avis d'expertise pour accompagner les experts en sûreté de fonctionnement des logiciels critiques 182

Session 10 : TALN, ontologie et graphe de connaissances 188

Ngoc Luyen Le, Marie-Hélène Abel et Philippe Gouspillou
Apport des ontologies pour le calcul de la similarité sémantique au sein d'un système de recommandation 189

Marion Schaeffer et Christophe Bouvard
Comparaison des solutions de NLU sur un corpus français pour un chatbot de support COVID-19

199

Nadia Yacoubi Ayadi, Catherine Faron, Franck Michel, Robert Bossy et Arnaud Barbe
Construction d'un graphe de connaissances à partir des annotations d'articles scientifiques et de leur contenu en sciences de la vie 209

Éditorial

Journées francophones d'Ingénierie des Connaissances

Les journées francophones d'Ingénierie des Connaissances (IC) sont organisées chaque année depuis 1997, d'abord sous l'égide du Gracq (Groupe de Recherche en Acquisition des Connaissances) puis sous celle du collège SIC (Science de l'Ingénierie des Connaissances) de l'AFIA. Cette année encore, IC est hébergée par la plateforme PFIA, conjointement avec d'autres conférences francophones dans le domaine de l'intelligence artificielle (IA).

L'ingénierie des connaissances peut être vue comme la thématique de l'Intelligence Artificielle accompagnant l'évolution des sciences et technologies de l'information et de la communication qui engendrent des mutations dans les pratiques individuelles et collectives. Elle ambitionne de contribuer à son essor en développant les modèles, les méthodes et les outils pour l'acquisition, la représentation et l'intégration de connaissances afin de rendre possible leur exploitation dans des environnements informatiques aux caractéritiques variées. La représentation formelle de ces connaissances permet des raisonnements automatiques sur ces connaissances et sur les données qui leur sont associées, pouvant être complexes, hétérogènes et évolutives. Sa finalité est la production de systèmes « intelligents et explicables », capables d'aider l'humain dans ses activités et pour la prise de décisions.

La conférence IC est un lieu d'échanges et de réflexions, de présentation et de confrontation des théories, pratiques, méthodes et outils autour de l'ingénierie des connaissances. Cette communauté prend désormais en compte l'essor des algorithmes d'apprentissage automatique et leurs retombées sur les pratiques individuelles et collectives, tout en conservant l'humain au centre des systèmes de décision exploitant les données et les connaissances.

Pour cette édition 2022 de la conférence, nous avons l'honneur de recevoir Prof. Christian Bizer - Chair of Information Systems V : Web-based Systems, Director of the Institute of Computer Science and Business Informatics, Allemagne, dont la conférence invitée est intitulée « *Integrating Product Data from the Semantic Web using Deep Learning Techniques* »

Suite à l'appel à contributions, la conférence IC a reçu 34 soumissions d'articles de travaux originaux et de travaux déjà publiés dans une conférence ou revue internationale de renom. Grâce au travail conséquent des membres du comité de programme, chaque article a reçu entre 3 et 4 relectures comportant des critiques argumentées et constructives pour les auteurs. Sur la base de ces critiques, le comité de programme qui s'est réuni en distanciel a sélectionné 10 articles longs et 8 articles courts de travaux originaux dans les thèmes de la conférence. Il a également sélectionné 11 articles de travaux déjà publiés et résumés en Français. Le programme de la conférence réparti sur 3 jours est organisé en 10 sessions dont le contenu est détaillé dans ces actes. Certaines sessions portent sur des thèmes qui sont au coeur de l'ingénierie des connaissances tels que « *la modélisation de connaissances complexes* », « *l'extraction d'informations et graphes de connaissances* » et « *le raffinement de graphes de connaissances* ». D'autres sessions concernent des thèmes émergeants dans la communauté tels que « *l'explicabilité et interprétabilité dans les graphes de connaissances* » et « *l'apprentissage automatique, ontologies et graphes de connaissances* »

Cette édition 2022 marque le retour de la conférence en présentiel après deux éditions de la plateforme PFIA qui se sont déroulées en distanciel en raison de la pandémie de Covid-19. Nous constatons une augmentation significative du nombre de soumissions et une implication très forte du comité de programme qui a oeuvré pour le succès de cette édition 2022 de la conférence IC.

Il nous reste à remercier chaleureusement l'ensemble des acteurs de la communauté francophone d'Ingénierie des Connaissances qui ont contribué au succès d'IC 2022, ainsi que le comité d'organisation de la plateforme PFIA 2022 qui a été d'une aide précieuse et qui nous a permis de nous réunir à nouveau en présentiel.

Fatiha Saïs

Comité de programme

Président

— Fatiha Saïs - Université Paris Saclay/LISN

Membres

— Marie-Hélène Abel - Université de technologie de Compiègne ;
— Xavier Aimé - Cogsonomy ;
— Yamine Ait-Ameur - INPT/IRIT ;
— Nathalie Aussenac-Gilles - CNRS/IRIT ;
— Bruno Bachimont - University de technologie de Compiègne ;
— Nacera Bennacer - CentraleSupélec ;
— Nathalie Bricon-Souf - Université Toulouse 3/IRIT ;
— Sandra Bringay - LIRMM, Université Paul Valéry ;
— Patrice Buche - INRAE ;
— Davide Buscaldi - École Polytechnique ;
— Sylvie Calabretto - INSA de Lyon ;
— Jérôme David - INRIA, Université Grenoble Alpes ;
— Pierre-Antoine Champin - ERCIM ;
— Jean Charlet - Assistance Publique hôpitaux de Paris ;
— Victor Charpenay - MINES Saint-Étienne ;
— Olivier Corby - Université Côte d'Azur ;
— Sylvie Despres - Université Sorbonne Paris Nord/LIMICS ;
— Gilles Falquet - University of Geneva, Switzerland ;
— Catherine Faron - Université Côte d'Azur ;
— Béatrice Fuchs - Université Lyon 3 ;
— Frederic Furst - Université de Picardie ;
— Alban Gaignard - CNRS ;
— Jean-Gabriel Ganascia - LIP6 ;
— Ollivier Haemmerlé - Université Toulouse 2/IRIT ;
— Mounira Harzallah - Université de Nantes ;
— Nathalie Hernandez - Université Toulouse 2/IRIT ;
— Liliana Ibanescu - Agro Paris Tech ;
— Sébastien Iksal - Le Mans Université ;
— Antoine Isaac - Europeana ;
— Clement Jonquet - INRAE ;
— Mouna Kamel - Université de Perpignan Via Domitia/IRIT ;
— Gilles Kassel - Université de Picardie Jules Verne ;
— Pascale Kuntz - Université de Nantes ;
— Michel Leclère - LIRMM ;
— Marie Lefèvre - Université Lyon 1 ;
— Dominique Lenne - Université de technologie de Compiègne ;
— Cedric Lopez - emvista ;
— Pascal Molli - Université de Nantes ;
— Isabelle Mougenot - Université de Montpellier ;
— Fleur Mougin - Université de Bordeaux ;
— Jérôme Nobécourt - Université Sorbonne Paris Nord/LIPN ;
— Nathalie Pernelle - Université Sorbonne Paris Nord/LIPN ;
— Yannick Prié - Université de Nantes ;
— Cedric Pruski - LIST, Luxembourg ;
— Sylvie Ranwez - IMT Mines Ales ;
— Catherine Roussey - INRAE ;
— Pascal Salembier - UTT ;
— Karim Sehaba - CNRS/LIRIS ;
— Danai Symeonidou - INRAE ;
— Rallou Thomopoulos - INRAE ;

— Cassia Trojahn - Université Toulouse 2/IRIT ;
— Raphael Troncy - Eurecom ;
— Haïfa Zargayouna - Université Sorbonne Paris Nord.

Session 1 : Extraction d'informations et graphes de connaissances (1)

DAGOBAH: Annotation sémantique de données tabulaires par comparaison du contexte des tables et d'un graphe de connaissances

Viet-Phi Huynh[1], Jixiong Liu[1,2], Yoan Chabot[1], Frédéric Deuzé[1], Thomas Labbé[1], Pierre Monnin[1], Raphaël Troncy[2]

[1] Orange, France
[2] EURECOM, Sophia Antipolis, France

Résumé

Cet article présente les améliorations apportées à DAGOBAH, un système effectuant un pré-traitement automatique et une interprétation sémantique de données tabulaires en fonction d'un graphe de connaissances. Nous détaillons les optimisations des mécanismes de recherche de candidats et les nouvelles techniques d'étude du contexte des nœuds du graphe de connaissances cible qui nous ont permis d'obtenir les meilleures performances lors du challenge SemTab 2021 en terme de précision. Nous décrivons également le déploiement des algorithmes DAGOBAH au sein de l'entreprise Orange via l'API TableAnnotation et une interface utilisateur. Ces deux méthodes d'accès permettent d'accélérer l'adoption de solutions d'interprétation de tables au sein de l'entreprise pour répondre à des besoins industriels.

Mots-clés

Interpretation sémantique de tables, DAGOBAH, SemTab

Abstract

In this paper, we present the latest improvements of the DAGOBAH system that performs automatic pre-processing and semantic interpretation of tables. In particular, we report promising results obtained in the SemTab 2021 challenge thanks to optimisations in lookup mechanisms and new techniques for studying the context of nodes in the target knowledge graph. We also present the deployment of DAGOBAH algorithms within the Orange company via the TableAnnotation API and a front-end DAGOBAH user interface. These two access methods enable to accelerate the adoption of Semantic Table Interpretation solutions within the company to meet industrial needs.

Keywords

Semantic Table Interpretation, DAGOBAH, SemTab

1 Introduction

Les données tabulaires constituent une source importante de connaissances, une grande partie des gisements internes des entreprises et du Web étant représentée sous cette forme. Par conséquent, il existe un vif interêt pour le domaine de l'interprétation automatique de données tabulaires (en anglais *Semantic Table Interpretation* ou STI). Ce domaine est caractérisé par le développement de méthodes d'interprétation automatique de tables à l'aide d'un graphe de connaissances via trois tâches principales. La tâche *Cell-Entity Annotation* (CEA) consiste à associer chaque cellule de la table avec une entité du graphe de connaissances. Par exemple, la mention "Belfort" de la Figure 1 sera annotée avec l'entité Q171545 (Belfort) du graphe de connaissances Wikidata. La tâche *Column-Type Annotation* (CTA) vise à annoter chaque colonne avec une classe. Par exemple, la première colonne "City" de la Figure 1 sera annotée avec l'entité Q484170 (commune française). Enfin, la tâche *Columns-Property Annotation* (CPA) vise à associer chaque paire de colonnes à une propriété. Par exemple, la relation entre les colonnes "City" et "Region" dans la Figure 1 serait la propriété P361 (fait partie de). Les annotations ainsi générées peuvent être utilisées dans de nombreux cas d'utilisation, de l'indexation des jeux de données et leur recommandation jusqu'à l'enrichissement de graphes de connaissances.

Les algorithmes de STI DAGOBAH, développés conjointement par Orange et EURECOM, ont été évalués lors des différentes éditions du challenge international SemTab [1] [8, 13, 14], colocalisé avec la conférence ISWC. Cet évènement centré sur les problématiques d'annotations de données tabulaires a attiré près de 50 équipes participantes au cours des trois dernières éditions. Comme démontré lors de ce challenge, nos outils ont atteint un niveau de maturité permettant de répondre à des problématiques industrielles dans le groupe. En effet, Orange est une multinationale opérant dans un grand nombre de domaines métiers (e.g. télécommunications, contenu multimédia, cybersécurité, etc.). Par conséquent, Orange produit de grands volumes de données tabulaires hétérogènes. A l'aide des techniques de STI, ces données peuvent être exploitées stratégiquement, par exemple, en structurant les connaissances dormantes dans ces données et en les rendant exploitables par le biais de moteurs de type questions-réponses [4].

Le challenge SemTab2021 et les besoins industriels mentionnés ci-dessus ont motivé des travaux de recherche qui constituent le cœur des algorithmes utilisés par le système DAGOBAH SL présenté en 2020 [11] et amélioré en

1. Semantic Web Challenge on Tabular Data to Knowledge Graph Matching : http://www.cs.ox.ac.uk/isg/challenges/sem-tab/

2021 [10]. En particulier, le calcul de scores et le classement des entités ont été optimisés grâce à :

— Une amélioration des stratégies d'indexation et de mise en correspondance des entités pour améliorer la qualité et la couverture de la recherche de candidats (i.e. *lookup*) ;
— Une meilleure représentation et désambiguïsation des entités en exploitant plus efficacement leurs contextes (i.e. voisinage) dans le graphe de connaissances ;
— Un algorithme de notation (i.e. *scoring*) des entités amélioré et plus flexible exploitant à la fois des informations locales et des informations globales à la table étudiée.

Ces nouvelles contributions ont donné naissance au système DAGOBAH SL 2021 décrit dans la Section 3. Nous présentons les résultats de l'évaluation menée dans le cadre du challenge dans la Section 4. La Section 5 introduit les efforts autour de l'utilisabilité des systèmes d'annotation avec en particulier la mise à disposition d'une API REST `TableAnnotation` ainsi qu'une interface utilisateur nommée DAGOBAH UI. Enfin, la Section 6 offre des éléments de réflexions autour du challenge SemTab et de l'adoption des outils de STI au sein des entreprises.

2 Etat de l'art

L'approche courante pour réaliser la tâche de CEA consiste à effectuer des opérations de recherche syntaxique (e.g. Levenshtein), d'alignement d'ontologies ou d'exploitation de plongements [15]. La désambiguïsation des entités candidates est ensuite traitée comme une tâche de classement des candidats, en utilisant des heuristiques, des algorithmes tels que PageRank [9] ou des modèles basés sur les graphes [12]. Les principales approches sur le typage de colonnes (CTA) infèrent des classes à partir des entités produites par la tâche CEA. Des heuristiques plus ou moins complexes construites autour du vote majoritaire sont utilisées [17]. Enfin, l'extraction de relations (CPA) est généralement réalisée par la recherche de paires d'éléments en colonnes, *i.e.* types et entités préalablement choisis [19].

Récemment, le challenge SemTab a permis d'accélérer le développement des approches de STI. Une majorité d'entre elles prennent la forme de systèmes basés sur la recherche de candidats dans DBpedia et Wikidata, le calcul d'une similarité syntaxique et des votes majoritaires [1, 5, 7]. MTab4Wikidata [18] adopte la correspondance floue et la "recherche à deux cellules" pour améliorer la prise en charge des fautes d'orthographe et des ambiguïtés dans le contenu des tableaux. Ce système a remporté le premier prix des défis SemTab 2019 et SemTab 2020.

3 Système DAGOBAH SL 2021

DAGOBAH est un processus de bout en bout annotant des tables relationnelles avec des éléments d'un graphe de connaissances tel que Wikidata. Ce processus se compose de quatre étapes exécutées en séquence tel qu'illustré dans la Figure 1. Etant donné une table relationnelle en entrée, l'étape de pré-traitement détermine un ensemble de métadonnées de la table ainsi que les cibles de l'annotation (Section 3.1). Le module de recherche de candidats collecte ensuite des entités candidates dans le graphe de connaissances pour chaque cellule cible de la table (Section 3.2). Le module de notation préliminaire évalue chacun de ces candidats afin de déterminer un score de confiance (Section 3.3). Les étapes suivantes visent à générer les annotations CTA ainsi que les annotations CPA (Section 3.4). Enfin, les annotations précédentes sont mises à contribution pour générer les annotations CEA (Section 3.4).

3.1 Pré-traitement des données tabulaires

Dans des cas d'utilisation réels, l'annotation des tables se révèle complexe en partie à cause de l'absence d'informations préalables sur leur structure et leur contenu. Ainsi, leur pré-traitement peut faciliter leur annotation. C'est pourquoi DAGOBAH propose des méthodes de pré-traitement visant à générer des métadonnées sur les tables via quatre tâches principales : la détection d'orientation, l'extraction d'en-têtes, l'identification de colonne clé [2] et le typage primitif des colonnes. Le typage primitif permet de détecter des entités nommées (e.g. localisation, organisation, personne), des littéraux avec unités (e.g. distance, vitesse, température) ou des littéraux divers (e.g. email, URL, adresse IP) [2].

3.2 Recherche d'entités candidates

L'étape de pré-traitement (et plus particulièrement le typage primitif) permet d'identifier les colonnes d'une table éligibles à l'étape de recherche d'entités candidates. Soit e_m une cellule d'une colonne éligible. L'étape de recherche d'entités candidates extrait un ensemble d'entités candidates pertinentes $\mathcal{E}_c(e_m)$ d'un graphe cible. Le service de recherche de candidats de DAGOBAH est basé sur ElasticSearch et supporte actuellement Wikidata et DBpedia pour lesquels des indexes ont été générés :

Entités Wikidata. Le service de recherche de candidats collecte les items et les propriétés ainsi que leurs labels et alias dans toutes les langues disponibles. Pour augmenter la couverture du service, les alias associés à chaque entité sont enrichis avec 11 propriétés supplémentaires telles que P2561 (name), P1705 (native label) ou P742 (pseudonym).

Entités DBpedia. Le service collecte les ressources en anglais ainsi que leurs labels dans toutes les langues disponibles. Pour augmenter la couverture, les labels sont enrichis avec les valeurs de 25 propriétés telles que abbreviation, birthName ou originalTitle. En complément, les labels et les alias de toutes les entités redirigées sont également inclus.

Nous faisons la moyenne des distances d'édition sur les caractères et sur les tokens [3] pour évaluer la similarité entre

2. Actuellement, seul l'identification d'une colonne clé unique est supportée par l'outil.
3. `https://github.com/seatgeek/thefuzz`

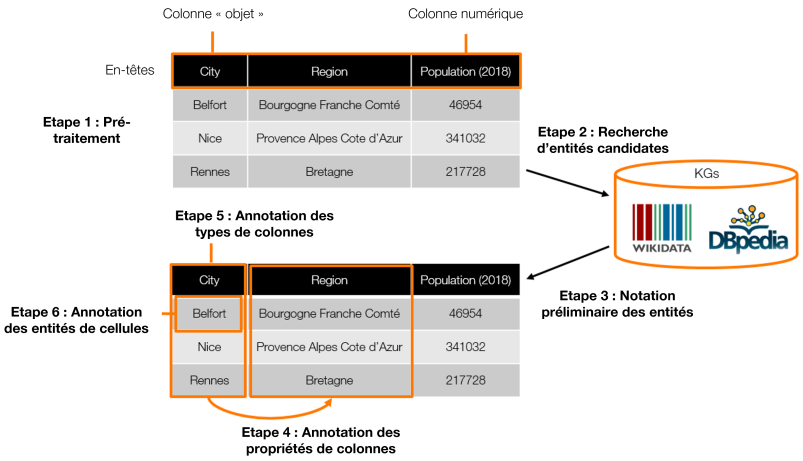


FIGURE 1 – Vue d'ensemble du processus d'annotation DAGOBAH.

une mention contenue dans une cellule et l'ensemble des labels de chaque entité candidate. Ce mode de fonctionnement permet de traiter au mieux les chaines de caractère présentant des ordonancements des sous-chaînes différents (e.g. "Elon Musk" et "Musk Elon").

3.3 Notation préliminaire des candidats

L'étape de notation préliminaire évalue la pertinence des entités candidates $e_c \in \mathcal{E}_c(e_m)$ d'une cellule e_m à l'aide d'un score :

$$PSc(e_c, e_m) = Sc_{context}(\mathcal{N}_{graph}(e_c), \mathcal{N}_{table}(e_m)) \times e^{\gamma(Sc_{sim}(e_c,e_m)-1)} \quad (1)$$

Ce score préliminaire est le produit d'un score de contexte et d'un score syntaxique $Sc_{sim}(e_c, e_m)$. Ce dernier facteur renvoie le plus haut ratio de correspondance, basé sur la distance de Levenshtein, entre la cellule et les labels et alias du candidat étudié. Les alias sont pénalisés avec un ratio pondéré par 0.9 car nous considérons que les labels ont plus d'importance pour la désambiguisation. Le facteur d'amplification $\gamma \in \mathbb{N}^+$ définit l'importance de la similarité syntaxique dans le calcul du score préliminaire. Nous avons déterminé, de manière empirique, que la valeur 2 était appropriée pour une utilisation du système sur les corpus du challenge SemTab2021.

Les améliorations du système DAGOBAH SL 2021 se concentre principalement sur le score de contexte, définit comme suit :

$$Sc_{context}(\mathcal{N}_{graph}(e_c), \mathcal{N}_{table}(e_m)) = \frac{\sum_i w_i \times sn_i}{\sum_i w_i} \quad (2)$$

où $\mathcal{N}_{table}(e_m)$ est l'ensemble des cellules voisines de e_m sur la même ligne et $\mathcal{N}_{graph}(e_c)$ est l'ensemble des nœuds voisins de l'entité e_c dans le graphe de connaissances [4]. Pour chaque cellule voisine $n_i \in \mathcal{N}_{table}(e_m)$, sn_i est un score de correspondance défini par rapport à $\mathcal{N}_{graph}(e_c)$.

4. Les nœuds voisins sont connectés à e_c via des chemins de prédicats dans le graphe de connaissances, quelque soit la direction des prédicats.

DAGOBAH SL 2021 résout deux problèmes de DAGOBAH SL 2020 inhérents au calcul du score de contexte :

Evaluation coûteuse. Chaque sn_i était évalué en itérant sur l'ensemble des nœuds dans $\mathcal{N}_{graph}(e_c)$ pour trouver la meilleure correspondance. Par conséquent, un problème de performance survient lorsque l'algorithme doit noter une entité très générique du graphe de connaissances présentant des centaines voire des milliers de propriétés. Par exemple, considérons la cellule "Belfort" dans la Figure 1 et l'entité Wikidata candidate Q171545. Pour vérifier si la cellule "Bourgogne Franche Comté" est dans le contexte de Q171545, nous devons effectuer une comparaison avec chacun des $\sim$ 1000 nœuds de $\mathcal{N}_{graph}(Q171545)$ ce qui inclus Q142 (France), Q3371185 (Paul Faivre), etc. (Figure 2a).

Contexte du graphe à un saut. $\mathcal{N}_{graph}(e_c)$ est l'ensemble des nœuds situés à un saut de e_c dans le graphe. Par conséquent, une cellule voisine $n_i \in \mathcal{N}_{table}(e_m)$ correspondant à un nœud situé à deux sauts de e_c n'était pas prise en compte dans le contexte de e_c. Par exemple, soit le contexte à un saut de Q171545 (Belfort) dans la Figure 2a, nous considérions, à tort, que Bourgogne Franche Comté n'avait pas de relations avec Belfort bien qu'il s'agisse de la région du Territoire de Belfort dont la capitale est Belfort.

DAGOBAH SL 2021 améliore l'efficacité du calcul et l'expressivité du score de contexte en évitant une notation exhaustive et en exploitant des contextes d'entités plus expressifs via la considération de nœuds à deux sauts.

3.3.1 Exploitation du contexte des entités du graphe de connaissances

Le score de correspondance du voisinage sn_i défini dans l'Equation (2) indique si une cellule voisine n_i de e_m correspond à un nœud voisin de e_c. Le calcul de sn_i peut se résumer à la recherche d'une entité candidate pour n_i

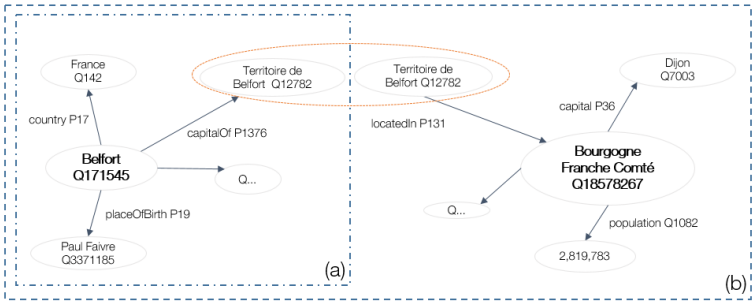


FIGURE 2 – Contexte (i.e. voisinage) de l'entité Q171545 (Belfort) dans le graphe Wikidata. (a) Contexte à un saut de Q171545. (b) Le contexte est étendu via l'intersection de sous-graphe.

dans $\mathcal{N}_{graph}(e_c)$ et à l'évaluation de sa similarité. Dans notre exemple précédent, Q18578267 est une entité candidate pour la cellule "Bourgogne Franche Comté" dans le contexte à deux sauts $\mathcal{N}_{graph}(Q171545)$ (Figure 2b). A partir de cette observation, nous proposons une méthode pour calculer efficacement le score sn_i. L'étape de recherche d'entités candidates (Section 3.2) génère des entités candidates $\mathcal{E}_c(e_m)$ pour une cellule cible e_m mais également des entités candidates $\mathcal{E}_c(n_i)$ pour des cellules voisines n_i. Par conséquent, nous vérifions si une entité candidate $e_i \in \mathcal{E}_c(n_i)$ est dans $\mathcal{N}_{graph}(e_c)$. Dans ce cas, sn_i est simplement calculé en comparant les labels de la cellule voisine n_i et le nœud correspondant e_i. Ce point permet d'éviter des comparaisons additionnelles avec d'autres nœuds de $\mathcal{N}_{graph}(e_c)$.

Pour vérifier si $e_i \in \mathcal{E}_c(n_i)$ est dans $\mathcal{N}_{graph}(e_c)$, nous vérifions si e_i est connecté à e_c via un chemin de prédicats dans le graphe de connaissances. Le calcul de ces chemins est un élément important dans le calcul du score. Pour trouver efficacement un chemin de prédicats entre e_c et e_i, nous extrayons les sous-graphes à un saut $\mathcal{G}_{e_c}$ et $\mathcal{G}_{e_i}$ de e_c et e_i. Si un nœud intermédiaire v est présent dans $\mathcal{G}_{e_c}$ et $\mathcal{G}_{e_i}$, les chemins pointant sur v dans les deux sous-graphes sont concaténés. Dans notre exemple, le chemin de prédicats suivant a été identifié : Belfort $\xrightarrow{\text{capitalOf}}$ Territoire de Belfort $\xrightarrow{\text{locatedIn}}$ Bourgogne Franche Comté. Seuls les sous-graphes à un saut étant pris en compte, les chemins de prédicats résultant ont une longueur maximum de deux. Cette approche permet d'enrichir les informations sur une entité en incluant non seulement les voisins directs mais également les voisins indirects à une distance de deux sauts. Ces contextes enrichis du graphe permettent d'augmenter les chances de correspondance avec une cellule voisine $n_i \in \mathcal{N}_{table}(e_m)$ et rendent ainsi le score de contexte plus précis. Après des tests, nous faisons l'hypothèse que pour l'interprétation de tables avec Wikidata, des chemins de taille supérieure à deux sont peu significatifs et apportent du bruit pouvant impacter négativement la pertinence du score de contexte.

3.3.2 Notation souple du contexte

Dans l'Equation (2), les scores de correspondances du voisinage sn_i sont pondérés pour calculer le score final d'une entité. En effet, chaque cellule voisine $n_i \in \mathcal{N}_{table}(e_m)$ contribue à un niveau différent à l'annotation de la cellule cible e_m avec un poids w_i défini par l'Equation (3) :

$$w_i = \underbrace{\frac{\overbrace{se_i}^{(3a)}}{\sqrt{d(col_i)+1}}}_{(3b)} \times \overbrace{cnt(col_i)}^{(3c)} \times \overbrace{\tau(e_i)}^{(3d)}. \qquad (3)$$

(3a) Les cellules contenant des entités devraient avoir une plus grande importance que les cellules contenant des littéraux (e.g. date, mesure avec ou sans unités, nombre) compte tenu du manque de méthodes de désambiguïsation des littéraux (e.g. normalisation des dates, détection des unités/normalisation/conversion). C'est la raison pour laquelle nous fixons la valeur de se_i à 1.0 dans le cas où la cellule voisine n_i contient une entité et à 0.15 si n_i contient un littéral.

(3b) Une cellule voisine sur la partie gauche de la table a plus de chance d'être un contexte significatif pour la cellule cible. Par conséquent, $d(col_i)$ est la distance entre la colonne col_i et la première colonne de type "entité" de la table.

(3c) Les cellules n_i appartenant à une colonne voisine très connectée à la colonne cible devrait avoir un plus grand poids dans le contexte. Par conséquent, nous prenons en compte la connectivité $cnt(col_i)$ d'une cellule voisine par rapport à la colonne cible. La connectivité est définie ici comme le nombre d'occurrence de la propriété la plus souvent observée entre les deux colonnes.

(3d) Les nœuds voisins de l'entité candidate e_c dans $\mathcal{N}_{graph}(e_c)$ peuvent fournir différents contenus informationnels étant donné que certains voisins peuvent être "sémantiquement plus proches" de e_c que d'autres. Par exemple, si nous considérons le contexte à deux sauts de l'entités Q171545 (Belfort) présenté dans la Figure 3, Q18578267 (Bourgogne Franche Comté) est plus pertinent que Q30 (United States of America) car le chemin Belfort $\xrightarrow{\text{capitalOf}}$ Territoire de Belfort $\xrightarrow{\text{locatedIn}}$

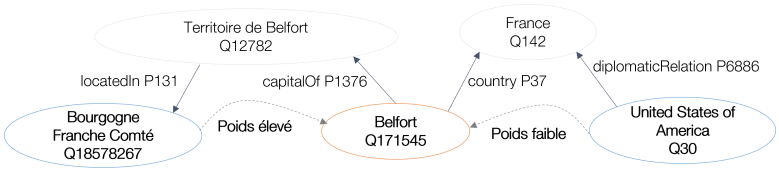


FIGURE 3 – Les nœuds voisins de Belfort (Q171545) contribuent de manière différente à son contenu informationnel.

Bourgogne Franche Comté porte davantage d'informations que le chemin Belfort $\xrightarrow{\text{country}}$ France $\xleftarrow{\text{diplomaticRelation}}$ United States of America. La valeur de vérité $\tau(e_i)$ [6] d'un nœud voisin e_i rend compte de cette différence en mesurant la capacité discriminative d'un chemin $\tau(e_c \xrightarrow{p_1} v \xrightarrow{p_2} e_i)$ et est défini comme suit :

$$\tau(e_i) = \tau(e_c \xrightarrow{p_1} v \xrightarrow{p_2} e_i) = \frac{1}{1 + log(g(v))} \quad (4)$$

où $g(v)$ est la généricité du nœud intermédiaire v, i.e. le nombre de ses relations entrantes et sortantes dans le graphe de connaissances. Il est à noter que les voisins directs (i.e. les chemins de longueur 1) obtiennent toujours une valeur de vérité de 1.0.

3.4 Tâches d'annotation

3.4.1 Annotation des propriétés de colonnes (CPA)

La tâche de CPA identifie la relation sémantique r la plus adaptée pour une paire ordonnée de colonnes. Nous adoptons une stratégie de vote majoritaire reposant sur les occurrences et les scores de confiances cumulés des lignes pour r (pour plus de détails, voir [11]). Il est à noter que, conformément à la Section 3.3, r peut être un chemin de prédicats de longueur 1 (i.e., $\xrightarrow{p}$), un chemin unidirectionnel de longueur 2 (i.e. $\xrightarrow{p1}\xrightarrow{p2}$ ou $\xleftarrow{p1}\xleftarrow{p2}$) ou un chemin bidirectionnel de longueur 2 (i.e. $\xrightarrow{p1}\xleftarrow{p2}$ ou $\xleftarrow{p1}\xrightarrow{p2}$).

3.4.2 Annotation des types de colonnes (CTA)

La tâche de CTA a pour but d'identifier le type le plus représentatif et le plus spécifique d'une colonne donnée. Pour cela, les types des entités candidates de la colonne sont collectés et une stratégie de vote majoritaire est appliquée pour déterminer le type le plus précis (voir [11] pour plus de détails sur les méthodes d'enrichissement de type et les calculs de scores).

3.4.3 Annotation des entités de cellules (CEA)

La tâche de CEA sélectionne pour une cellule e_m l'entité la plus pertinente parmi les entités candidates $e_c \in \mathcal{E}_c(e_m)$ collectées dans le graphe de connaissances. Cette étape s'appuie à la fois sur la notation préliminaire des entités et sur les informations fournies par le CTA et le CPA pour calculer la note finale des entités candidates. En effet, la notation préliminaire d'une entité candidate e_c tient uniquement compte des informations locales, i.e. les informations de la ligne à laquelle elle appartient. La prise en compte du type de colonne fourni par le CTA et de la propriété identifiée par le CPA permet de prendre en compte des informations globales. Par conséquent, le score final $Sc(e_c, e_m)$ d'une entité candidate e_c est calculé comme suit :

$$Sc(e_c, e_m) = \frac{(PSc(e_c, e_m) + \alpha \times score_{CTA} + \beta \times \overline{score_{CPA}})}{1 + \alpha + \beta} \quad (5)$$

Si e_c appartient au type généré par le CTA pour la colonne, alors $score_{CTA}$ est égal au score attribué à ce type et 0 dans le cas contraire. Via $\overline{score_{CPA}}$, nous calculons la moyenne des scores des relations identifiées par le CPA impliquant la colonne de e_c. Pour chaque relation, si e_c appartient au domaine ou au co-domaine (selon l'orientation de la relation), alors nous considérons le score de cette relation, sinon, le score est fixé à 0. Pour renforcer (resp. affaiblir) un CTA/CPA fréquent (resp. peu fréquent) lors de la mise à jour de $Sc(e_c, e_m)$, un coefficient α (resp. β) est utilisé et défini par $\frac{occurrence(CTA)}{2}$ (resp. $\frac{\overline{occurrence(CPA)}}{2}$). Il est à noter que le nombre d'occurrences du CTA/CPA est divisé par 2 pour accorder davantage d'importance au score préliminaire $PSc(e_c, e_m)$.

4 Evaluation

4.1 Configurations

Afin d'évaluer l'apport des contextes de graphe à un et à deux sauts ainsi que de la notation de contexte souple définis dans la Section 3, nous définissons quatre configurations pour les expériences :

Configuration 1 Le score de contexte d'une entité est calculé en utilisant uniquement le voisinage à un saut du graphe de connaissances. Les poids w_i ne sont pas calculés à l'aide de l'Equation (3) mais fixés à 1.0 pour les entités et 0.15 pour les littéraux.

Configuration 2 Le score de contexte d'une entité est calculé en utilisant le voisinage à deux sauts du graphe de connaissances. Les poids w_i ne sont pas calculés à l'aide de l'Equation (3) mais sont fixés à 1.0 pour les voisins à un saut, 0.25 pour les voisins à deux sauts et 0.15 pour les littéraux.

Configuration 3 Le score de contexte d'une entité est calculé en utilisant le voisinage à deux sauts du graphe de connaissances. Les poids w_i sont calculés à l'aide de l'Equation (3). Cette configuration permet de tester si des contextes plus riches associés à une notation stricte permet de générer de meilleures annotations.

Configuration 4 Ce paramétrage restreint la configuration 3 au voisinage à un saut et aux voisins liés par un chemin unidirectionnel de longueur 2 dans le graphe. Cette configuration permet d'évaluer l'impact des chemins bidirectionnels qui semblent être moins informatifs (et amenant parfois du bruit) mais utiles dans certains cas bien ciblés.

4.2 Résultats

4.2.1 Evaluation expérimentale

Les résultats pour les quatre configurations définies précédemment sont donnés dans la Table 1. Il est à noter que les performances de DAGOBAH se sont continuellement améliorées tout au long du challenge SemTab2021. Ainsi, les résultats de l'évaluation sont basés sur la dernière version de DAGOBAH mais nous indiquons également les résultats soumis lors des différentes phases du challenge dans les cellules grisées ainsi que le meilleur score parmi les participants du challenge [5], pour comparaison. Afin de valider la pertinence des modifications proposées dans les Sections 3.2 et 3.3, nous incluons également les scores du système DAGOBAH 2020 pour les tables du Round 1 annotées à l'aide de Wikidata. Les configurations des soumissions {1,2,3,4}* sont similaires aux configurations {1,2,3,4} définies précédemment avec quelques adaptations sur l'initialisation des scores et des poids. Cela n'impacte toutefois pas les scores de CEA mais a en revanche un impact sur les performances du CTA. En effet, le CTA est très sensible aux scores d'entités et aux poids attribués à la taxonomie pour la sélection du type le plus spécifique parmi l'ensemble des types possibles pour les entités (types directs, parents, etc.). DAGOBAH obtient d'excellents résultats sur les jeux de données synthétiques (Round 2) tandis que les jeux de données générés manuellement et présentant des dispositions plus complexes semblent être traités de manière moins satisfaisante (Rounds 1 et 3). Sur le corpus HardTable, l'utilisation de contextes plus riches et de la technique de notation souple ne semble pas amener de gain. Cela peut s'expliquer par le fait que les tableaux de ce corpus sont presque entièrement représentés dans le graphe de connaissances et que les colonnes peuvent donc être désambiguisées seulement à partir de leur contenu. A l'inverse, le corpus BioTable contient des ambiguités plus complexes avec des chevauchements de contenu entre les colonnes empêchant leur désambiguisation (e.g. la colonne "Gene" peut être confondue avec la colonne "Protein", les valeurs étant souvent similaires). L'annotation semble donc bénéficier de contextes de graphes plus riches. Pour BioDivTable, la configuration 4 est celle obtenant les scores les plus bas, tandis que la configuration 1 est comparable à la configuration 3. Nous supposons que les chemins unidirectionnels de longueur 2 apportent du bruit pouvant expliquer les faibles performances de la configuration 4.

En règle général, les configurations 2, 3 et 4 sont plus précises pour le CEA que la configuration 1. La récupération du contexte de graphe à deux sauts semble donc être un ajout bénéfique permettant de récupérer des informations pertinentes. Les meilleures performances des configurations 3 et 4 vis à vis de la configuration 2 montre l'efficacité de la notation de contexte souple. Nous notons que la configuration 3 atteint des performances proches de la configuration 4. Ainsi, les chemins unidirectionnels (i.e. $\xrightarrow{p_1}\xrightarrow{p_2}$ et $\xleftarrow{p_1}\xleftarrow{p_2}$) apportent suffisamment d'informations et permettent d'obtenir des résultats équivalents par rapport à la configuration considérant à la fois les chemins unidirectionnels et bidirectionnels. De plus, l'influence négative du bruit apporté par les chemins bidirectionnels (e.g. Belfort $\xrightarrow{\text{country}}$ France $\xleftarrow{\text{diplomaticRelation}}$ United States of America) est limitée par le calcul de score de contexte souple qui évite une dégradation de la qualité de l'annotation. Cela permet aux chemins bidirectionnels pertinents de contribuer positivement au score de l'entité. On peut observer que les performances du CTA et du CPA ne sont pas aussi élevées qu'envisagé sur la plupart des corpus, et ce, malgré de bonnes performances de CEA. Le développement de stratégies plus performantes pour la sélection du type et des relations fera l'objet de travaux futurs.

4.2.2 Corpus BioDivTab et GitTables

Il est à noter que pour les corpus BioDivTab et GitTables, nous avons adapté les algorithmes DAGOBAH présentés dans cet article. En effet, pour le corpus BioDivTab, les types primitifs générés par le pré-traitement ont été utilisés pour discriminer les colonnes "entités" et les colonnes contenant des littéraux. Une colonne contient des littéraux si elle contient des valeurs numériques, des dates, des unités ou des valeurs diverses. Sinon, la colonne est considérée comme une colonne d'entités et ses mentions peuvent donc être utilisées par le module de recherche de candidats. Pour le corpus GitTables, des règles de correspondance ont été définies entre les types primitifs et des classes de Schema.org et de l'ontologie DBpedia.

5 L'interprétation de données tabulaires à Orange

Pour améliorer la pertinence de DAGOBAH sur des cas d'utilisation industriels réels, nous avons adopté une approche Test & Learn. Dans cette optique, les algorithmes de DAGOBAH sont mis à disposition au sein de l'entreprise pour permettre aux collaborateurs internes de tester les outils d'annotation. Cette mise à disposition s'effectue via deux vecteurs : une API REST nommée `TableAnnotation` et une interface graphique nommée DAGOBAH UI.

5.1 API TableAnnotation

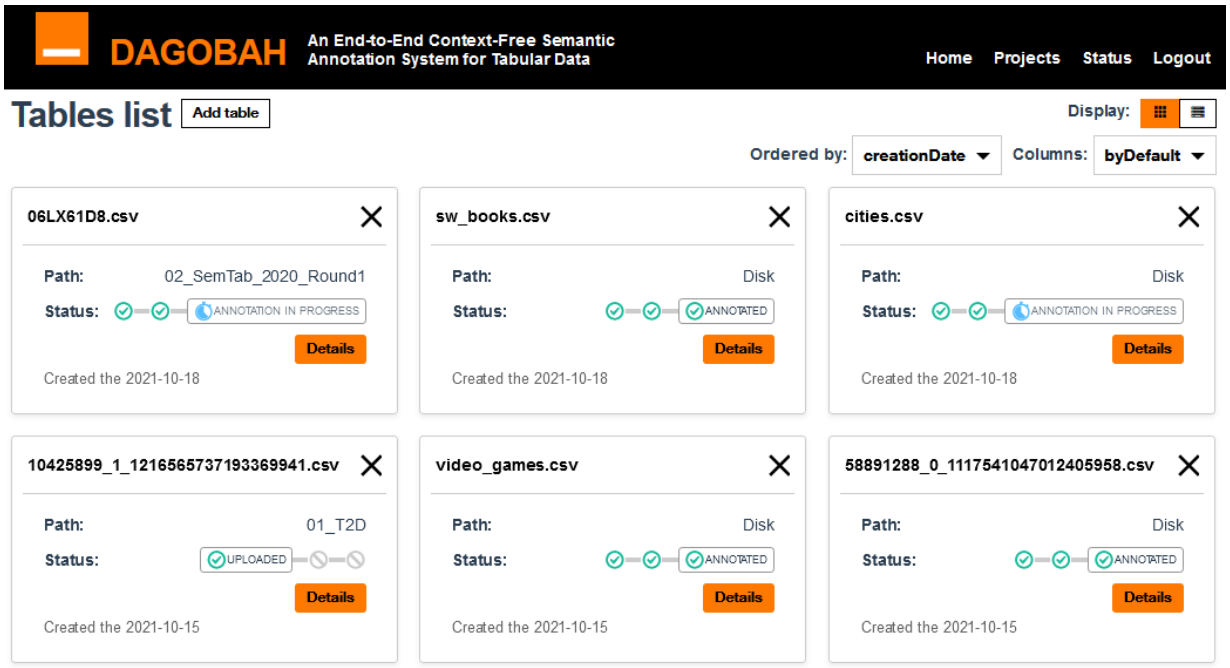
Cette API REST est déployée sur le portail Orange Developer. [6] Elle fournit des services de pré-traitement de données tabulaires, d'annotation sémantique et également de recherche d'entités candidates permettant, à partir d'une

5. Les résultats complets sont disponibles en ligne : `http://www.cs.ox.ac.uk/isg/challenges/sem-tab/2021/index.html#results`

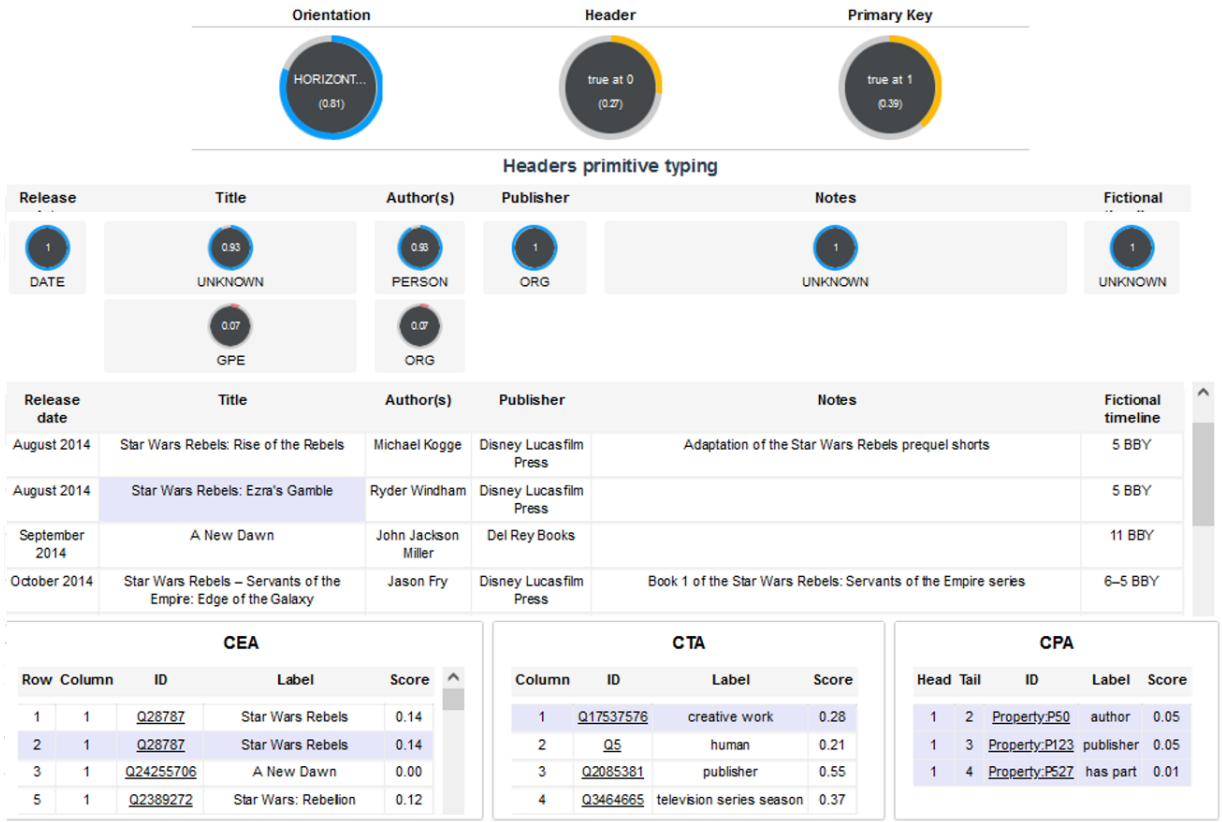
6. `https://developer.orange.com`

TABLE 1 – Comparaison des configurations expérimentales et des performances du système DAGOBAH sur les Rounds 1, 2, et 3 du challenge SemTab2021 (dans les cellules grisées). "F1" signifie F1-score, "P" signifie Precision. Les meilleurs résultats sont mis en valeur en gras.

Corpus	Configuration	CTA		CEA		CPA	
		F1	P	F1	P	F1	P
Round 1 – WDTable	Configuration 1	0.793	0.793	0.913	0.913	-	-
	Configuration 2	0.790	0.790	0.923	0.923	-	-
	Configuration 3	0.783	0.783	**0.926**	**0.926**	-	-
	Configuration 4	0.783	0.783	0.924	0.924	-	-
	DAGOBAH 2020	0.743	0.743	0.830	0.841	-	-
	Configuration 2*	**0.832**	**0.832**	0.923	0.923	-	-
	Top 1 SemTab2021	0.728	0.73	0.907	0.907	-	-
Round 1 – DBPTable	Configuration 1	0.25	0.25	0.935	0.935	-	-
	Configuration 2	0.27	0.27	0.946	0.946	-	-
	Configuration 3	0.274	0.274	**0.947**	**0.947**	-	-
	Configuration 4	0.274	0.274	**0.947**	**0.947**	-	-
	Configuration 2*	0.422	0.424	0.945	0.946	-	-
	Top 1 SemTab2021	**0.46**	**0.468**	0.692	0.692	-	-
Round 2 – BioTable	Configuration 1	0.874	0.874	0.882	0.882	0.898	**0.901**
	Configuration 2	0.911	0.911	0.916	0.916	**0.899**	0.899
	Configuration 3	0.915	0.915	0.950	0.951	**0.899**	0.899
	Configuration 4	0.916	0.916	**0.970**	**0.970**	**0.899**	0.899
	Configuration 4*	0.916	0.916	0.970	0.970	0.899	0.899
	Top 1 SemTab2021	**0.956**	**1**	0.964	0.964	0.899	0.899
Round 2 – HardTable	Configuration 1	0.968	0.969	0.975	0.976	0.996	0.997
	Configuration 2	0.968	0.969	0.976	0.976	0.996	0.997
	Configuration 3	0.968	0.969	0.976	0.976	0.996	0.997
	Configuration 4	0.968	0.968	0.976	0.976	0.996	0.997
	Configuration 3*	0.976	0.976	0.975	0.976	0.996	0.996
	Top 1 SemTab2021	**0.977**	**0.977**	**0.985**	**0.985**	**0.997**	**0.998**
Round 3 – BioDivTable	Configuration 1	0.338	0.339	0.619	0.64	-	-
	Configuration 2	0.335	0.335	0.60	0.62	-	-
	Configuration 3	0.344	0.345	**0.62**	**0.641**	-	-
	Configuration 4	0.343	0.343	0.475	0.491	-	-
	Configuration 4*	0.381	0.382	0.496	0.497	-	-
	Top 1 SemTab2021	**0.593**	**0.595**	0.602	0.611	0.947	1
Round 3 – HardTable	Configuration 3*	**0.99**	**0.99**	**0.974**	**0.974**	0.991	**0.995**
	Top 1 SemTab2021	0.984	0.984	0.968	0.968	**0.993**	0.994
Round 3 – GitTables DBP	Pré-traitement + Mapping	**0.07**	**0.117**	-	-		-
	Top 1 SemTab2021	0.041	0.042	-	-	-	-
Round 3 – GitTables SCH	Pré-traitement + Mapping	0.183	0.185	-	-	-	-
	Top 1 SemTab2021	**0.205**	**0.943**	-	-	-	-



(a) DAGOBAH UI permet de créer des projets et de charger des tables à partir du système de fichier local ou à partir de corpus de référence (e.g. T2D, SemTab).



(b) DAGOBAH UI permet d'afficher les résultats générés par les outils de pré-traitement et d'annotation. En partie haute, l'outil affiche les informations de pré-traitement (e.g. orientation, en-têtes) ainsi que la table nettoyée. En partie basse, une vue interactive permet à l'utilisateur de naviguer dans les annotations CEA, CTA et CPA.

FIGURE 4 – Fonctionnalités de DAGOBAH UI.

mention, de collecter des entités Wikidata ou DBpedia potentiellement correspondantes. Cette API est accessible à l'ensemble des collaborateurs des entités R&D du groupe ainsi que des unités d'affaires, sur invitation. Nous planifitions d'ouvrir plus largement l'accès à cette API dans un futur proche.

5.2 DAGOBAH UI

Cette interface graphique permet à des collaborateurs non familiers avec le développement ou l'intelligence artificielle d'utiliser les fonctions de l'API `TableAnnotation` sur leurs tables et de visualiser les résultats sous une forme intelligible et ergonomique. Les utilisateurs ont la possibilité de charger des tables dans leurs projets d'annotation (Figure 4a) puis de lancer le pré-traitement de ces dernières ainsi que l'annotation sémantique. Les résultats de ces processus peuvent ensuite être visualisés (Figure 4b). DAGOBAH UI est un outil très puissant pour démontrer la valeur des techniques d'interprétation automatique de tables au sein du groupe Orange mais également auprès de prospects externes. Une vidéo de démonstration de l'interface graphique est disponible à `https://tinyurl.com/dagobah-ui`. Les développements récents sur cette interface permettent *(i)* l'enrichissement de graphes de connaissances à partir d'éléments de la table non présents dans le graphe, *(ii)* l'enrichissement de la table à partir du graphe de connaissances afin de compléter des valeurs manquantes ou d'ajouter de nouvelles colonnes, et *(iii)* la visualisation interactive du graphe de connaissances cible avec une mise en valeur des annotations résultants des étapes de CEA, CTA, CPA ainsi que des nouveaux triplets générés à partir de la table.
Cette interface utilisateur permet aux collaborateurs de saisir l'interêt de l'annotation sémantique pour des cas d'utilisation industriels. Inversement, l'équipe de recherche DAGOBAH peut identifier les défis associés à ces cas d'utilisation, ce qui constitue un apport précieux pour la feuille de route du projet. Bien que le déploiement et l'adoption des méthodes de STI chez Orange n'en soient qu'à leurs débuts, des tests sur différents cas d'utilisation ont lieu depuis plus d'un an via l'API `TableAnnotation` qui a répondu à plus de 200 000 requêtes. Plusieurs domaines sont envisagés comme cibles prioritaires pour l'annotation sémantique, incluant le divertissement (e.g. annotation de catalogues de films), la gouvernance des données ou la santé.

6 Discussion

Les corpus de données proposés cette année par le challenge SemTab2021 ont permis de prendre en compte une plus grande variété de problématiques associées à l'interprétation automatique de données tabulaires. Cette édition a notamment intégré de nouveaux domaines de connaissances (e.g. biomédical et données Git) et a ajouté de nouvelles contraintes sur l'annotation avec le support de graphes de connaissances multiples (Wikidata et DBpedia) et l'annotation à l'aide de schémas uniquement (Schema.org et l'ontologie DBpedia). Ces challenges nous ont permis d'améliorer les stratégies d'annotation du système DAGOBAH avec notamment l'exploitation des types primitifs générés par le pré-traitement et l'utilisation de contextes de graphes enrichis.

Néanmoins, de nouvelles directions de recherche peuvent encore être explorées pour faire face à l'hétérogénéité des types de tableaux publiés sur le Web. Ainsi, la structure des tableaux et les relations internes pourraient être prises en compte (e.g. orientation des tableaux, cellules imbriquées, concaténation pour mise en page, cellules à valeurs multiples, sujets répartis dans plusieurs colonnes comme les noms et prénoms d'une personne, etc). De plus, une problématique demeure dans le traitement des données hors graphe de connaissances, i.e. des entités non présentes dans un graphe cible donné, ce qui est souvent le cas pour des données spécifiques aux entreprises. Il convient de noter que les données hors graphe de connaissances ont commencé à être abordées avec le corpus GitTables. Ce dernier nécessitait en effet d'annoter des tables avec Schema.org et l'ontologie DBpedia uniquement. Cependant, cette tâche n'était pas entièrement conforme avec la définition du CTA utilisée par la communauté car les annotations recherchées mélangeaient des classes et des propriétés. Ces annotations hétérogènes peuvent conduire à des évaluations incohérentes. Au delà des données hors graphe, il serait intéressant d'évaluer la portabilité de l'approche à des graphes de domaines (e.g. biomédicaux, linguistiques) et leur apport pour l'annotation de jeux de données spécifiques.

Nous avons récemment proposé une classification reflétant l'hétérogénéité des tables que l'on peut rencontrer ainsi qu'un inventaire exhaustif des méthodes, à base de règles et d'heuristiques, ou à base d'apprentissage profond pour l'interprétation sémantique de données tabulaires [16]. Nos travaux en cours se concentrent sur l'interprétation de tables où une grande part des mentions d'une colonne ne trouve pas de correspondance, ou avec peu de lignes et donc peu de contexte. Dans ces cas difficiles, nous cherchons à tirer profit des modèles de langage et des méthodes de plongement de graphes qui pourraient apporter un complément intéressant aux stratégies de calcul de score de contexte.

7 Conclusion

Dans cet article, nous avons présenté les améliorations apportées au système DAGOBAH [3]. Grace à un mécanisme de recherche de candidats optimisé, l'enrichissement des contextes du graphe et la notation souple, DAGOBAH a obtenu la meilleure performance lors du challenge SemTab2021.[7] Les travaux futurs auront pour objectif d'augmenter la précision de l'annotation sur des tables présentant des mentions très ambiguës. Nous avons notamment l'ambition d'exploiter des dictionnaires fournissant des abréviations ou des acronymes. Pour assurer la généricité de notre approche, de tels dictionnaires devraient être construits à partir de grandes quantités de documents et être applicables à divers ensembles de données.

7. `https://www.cs.ox.ac.uk/isg/challenges/sem-tab/2021/index.html#results`

Remerciements

Les auteurs remercient Christophe Sarthou-Camy et Guillaume Jourdain pour leurs contributions importantes dans le développement de DAGOBAH UI.

Références

[1] Nora Abdelmageed and Sirko Schindler. JenTab Meets SemTab 2021's New Challenges. In *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab). CEURWS. org*, 2021.

[2] Yoan Chabot, Thomas Labbé, Jixiong Liu, and Raphaël Troncy. DAGOBAH : An End-to-End Context-Free Tabular Data Semantic Annotation System. In *International Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*, volume 2553 of *CEUR Workshop Proceedings*, pages 41–48, 2019.

[3] Yoan Chabot, Thomas Labbé, Jixiong Liu, and Raphaël Troncy. DAGOBAH : Un système d'annotation sémantique de données tabulaires indépendant du contexte. In *31es Journées francophones d'Ingénierie des Connaissances (IC)*, Angers, France, 2020.

[4] Yoan Chabot, Pierre Monnin, Frédéric Deuzé, Viet-Phi Huynh, Thomas Labbé, Jixiong Liu, and Raphaël Troncy. A Framework for Automatically Interpreting Tabular Data at Orange. In *20th International Semantic Web Conference (ISWC), Posters, Demos and Industry Tracks*, volume 2980 of *CEUR Workshop Proceedings*, 2021.

[5] Shuang Chen, Alperen Karaoglu, Carina Negreanu, Tingting Ma, Jin-Ge Yao, Jack Williams, Andy Gordon, and Chin-Yew Lin. Linkingpark : An integrated approach for semantic table interpretation. In *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*, 2020.

[6] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. Computational fact checking from knowledge networks. *PloS one*, 10(6), 2015.

[7] Marco Cremaschi, Roberto Avogadro, Andrea Barazzetti, and David Chieregato. MantisTable SE : an Efficient Approach for the Semantic Table Interpretation. In *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*, 2020.

[8] Vincenzo Cutrona, Jiaoyan Chen, Vasilis Efthymiou, Oktie Hassanzadeh, Ernesto Jiménez-Ruiz, Juan Sequeda, Kavitha Srinivas, Nora Abdelmageed, Madelon Hulsebos, Daniela Oliveira10, et al. Results of SemTab 2021. In *CEUR Workshop Proceedings*, 2021.

[9] Vasilis Efthymiou, Oktie Hassanzadeh, Mariano Rodriguez-Muro, and Vassilis Christophides. Matching web tables with knowledge base entities : From entity lookups to entity embeddings. In *16th International Semantic Web Conference (ISWC)*, pages 260–277, 2017.

[10] Viet-Phi Huynh, Jixiong Liu, Yoan Chabot, Frédéric Deuzé, Thomas Labbé, Pierre Monnin, and Raphaël Troncy. DAGOBAH : Table and Graph Contexts For Efficient Semantic Annotation Of Tabular Data. In *International Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*, volume 3103 of *CEUR Workshop Proceedings*, pages 19–31, 2021.

[11] Viet-Phi Huynh, Jixiong Liu, Yoan Chabot, Thomas Labbé, Pierre Monnin, and Raphaël Troncy. DAGOBAH : enhanced scoring algorithms for scalable annotations of tabular data. In *International Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*, volume 2775 of *CEUR Workshop Proceedings*, pages 27–39, 2020.

[12] Yusra Ibrahim, Mirek Riedewald, and Gerhard Weikum. Making sense of entities and quantities in Web tables. In *International Conference on Information and Knowledge Management (CIKM)*, pages 1703–1712, 2016.

[13] Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, Vasilis Efthymiou, Jiaoyan Chen, and Kavitha Srinivas. SemTab 2019 : Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems. In *European Semantic Web Conference (ESWC)*, pages 514–530. Springer, 2020.

[14] Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, Vasilis Efthymiou, Jiaoyan Chen, Kavitha Srinivas, and Vincenzo Cutrona. Results of SemTab 2020. In *CEUR Workshop Proceedings*, volume 2775, pages 1–8, 2020.

[15] Girija Limaye, Sunita Sarawagi, and Soumen Chakrabarti. Annotating and searching web tables using entities, types and relationships. In *36th International Conference on Very Large Data Bases (VLDB)*, pages 1338–1347, 2010.

[16] Jixiong Liu, Yoan Chabot, Raphaël Troncy, Viet-Phi Huynh, Thomas Labbé, and Pierre Monnin. From Tabular Data to Knowledge Graphs : A Survey of Semantic Table Interpretation Tasks and Methods. *To appear in the Journal of Web Semantics*, 2022.

[17] Varish Mulwad, Tim Finin, Zareen Syed, and Anupam Joshi. Using linked data to interpret tables. In *1st International Workshop on Consuming Linked Data (COLD)*, 2010.

[18] Phuc Nguyen, Ikuya Yamada, Natthawut Kertkeidkachorn, Ryutaro Ichise, and Hideaki Takeda. SemTab 2021 : Tabular Data Annotation with MTab Tool. In *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*, 2021.

[19] Chenwei Ran, Wei Shen, Jianyong Wang, and Xuan Zhu. Domain-specific knowledge base enrichment using wikipedia tables. In *IEEE International Conference on Data Mining (ICDM)*, pages 349–358, 2016.

Connaissances géospatiales dans les annonces immobilières : détection et extraction d'information spatiale à partir du texte

Lucie Cadorel[1,3], Alicia Blanchi[2,3], Andrea G. B. Tettamanzi[1]

[1] Université Côte d'Azur, Inria, CNRS, I3S
[2] Université Côte d'Azur, ESPACE, CNRS
[3] KCityLabs

lucie.cadorel@inria.fr

Résumé

Nous avons proposé un modèle d'extraction de connaissances géospatiales à parir du texte appliqué au cas des annonces immobilières. La première étape consiste à extraire les entités géographiques et spatiales à l'aide d'un modèle basé sur une architecture BiLSTM-CRF *et la concaténation de plusieurs embeddings. Ensuite, nous avons réalisé l'extraction de relations, notamment spatiales, pour créer une base de connaissance géospatiale structurée stockée dans un graphe de connaissance RDF.*

Mots-clés

Extraction d'information, connaissance géographique, reconnaissance d'entités nommées, extraction de relation

Abstract

We proposed a workflow to extract geospatial knowledge from text applied to Real Estate advertisements. We first extracted geographic and spatial entities using a model based on a BiLSTM-CRF *architecture with a concatenation of several text representations. Secondly, we performed relations extraction, particularly spatial relations extraction, to build a structured Geospatial knowledge base that we stored in a RDF Knowledge Graph.*

Keywords

Information extraction, geographical knowledge, named entity recognition, relationship extraction

1 Introduction

La reconnaissance d'entités géospatiales dans les textes a largement été développée par les avancées en traitement du langage naturel, et a été appliquée à divers types de textes tels que les blogs de voyage [2], les réseaux sociaux [3] ou bien les annonces immobilières [4]. L'approche traditionnelle consiste à utiliser des règles linguistiques et des dictionnaires géographiques (gazetteer). Cependant, cette approche donne des résultats limités puisqu'elle dépend de la complétude des règles et des dictionnaires. Ainsi, les modèles utilisant du Deep Learning sont de plus en plus développées et obtiennent de très bons résultats. Néanmoins, la plupart des études detectent seulement les lieux-nommées alors que des termes géographiques (e.g., la gare, la plage, les écoles, etc.) peuvent être aussi utilisés pour mentionner un lieu. De plus, les relations spatiales sont aussi source d'information et permettent de mieux localiser un lieu (e.g, "à 10 minutes", "proche", "à deux pas", etc.) mais sont rarement extraites. On retrouve notamment ce type de connaissances dans les annonces immobilières. En effet, les agents immobiliers décrivent de façon vague les lieux qui permettent de situer une propriété (e.g, "L'appartement est situé dans un quartier résidentiel proche de l'université de Nice. Commodités à deux pas."). Cependant, ces lieux flous donnent des informations à la fois sur la localisation du logement et sur le quartier et ses équipements. Il est donc important de reconnaître et d'extraires ces connaissances afin de mieux comprendre le marché immobilier, les éléments influents les prix ou encore la perception des lieux résidentiels . Ainsi, les agents immobiliers pourront mieux connaître les prix, les tendances du marché d'un quartier et les biens similairs à celui vendu, notamment lorsque celui-ci est situé hors du secteur habituel de l'agent.

Nous avons donc proposé un modèle d'extraction de connaissances géospatiales à parir du texte appliqué au cas des annonces immobilières. Cet article est un résumé traduit et mis à jour de l'article [1] que nous avons publié à K-CAP '21.

Le reste de cet article est organisé de la manière suivante. Dans la section 2, nous présentons le pipeline d'extraction mis en place pour retrouver les entités géospatiales et les relations puis les stocker de manière structurée. La section 3 détaille l'évaluation et la comparaison du modèle proposé.

2 Extraction d'information géospatiale

2.1 Reconnaissance d'entités géospatiales

La reconnaissance d'entités nommées géospatiales consiste à identifier les termes d'un texte faisant référence à des entités géographiques et spatiales telles que les lieux-nommés ("Nice", "Place Masséna", etc.). Nous avons identifié quatre catégories à extraire : Lieu-nommé (Toponym), type de lieu (Feature), entité spatio-temporelle (Spatiotemporal) et le mode de transport (Mode of transporation). Les deux pre-

mières catégories définissent explicitement un lieu à différents niveaux de précision, tandis que les entités spatio-temporelles et le mode de transport décrivent une relation spatiale permettant de localiser ce lieu.

Pour extraire ces informations, nous avons mis au point un modèle basé sur une architecture *BiLSTM+CRF* prenant en entrée un embedding du texte. Nous l'avons entraîné sur un corpus d'environ 1200 annonces immobilières préalablement annotées en utilisant le format de tag BIESO. L'embedding utilisé est un vecteur composé de la concaténation de trois représentations différentes du texte. La première représentation est un Word Embedding classique entraîné sur notre corpus d'annonces immobilières. La seconde est basé sur le modèle de langage pré-entraîné Flair pour le français que nous avons réentraîné sur notre corpus. Enfin, nous utilisons le modèle de langage CamemBERT sans réentraînement dû au manque d'un corpus de taille suffisante. Ces trois représentations permettent de capturer les spécificités et la variabilité du style de langage utilisé dans les annonces immobilières.

2.2 Extraction de relations

La deuxième partie de notre travail vise à obtenir une représentation structurée des informations extraites. Pour cela, nous avons extrait trois types de relations entre les entités retrouvées : Attribut, Type de lieu nommé et Spatiale.

Nous avons fait plusieurs hypothèses pour extraire les relations. D'abord, une relation a lieu seulement entre deux entités d'une même phrase. Il existe donc toujours un lien direct ou indirect entre les deux entités qui peut être ainsi retrouvé à l'aide d'un graphe de dépendance grammaticale. Pour obtenir ce graphe de dépendance, nous utilisons un modèle d'analyse de dépendance qui renvoie la structure syntaxique d'une phrase à partir de la grammaire. Ce modèle détermine les connections grammaticales entre les mots suivant le schéma *<Sujet, Fonction grammaticale, Objet>* qui est adapté à la structure syntaxique des annonces immobilières. En effet, celles-ci ne suivent pas toujours la grammaire standard avec un ordre des mots différents, un sujet ou un verbe absent, etc. Le modèle utlisé est l'analyseur syntaxique de Stanza pour le français basé sur la taxonomie universelle des dépendances (Universal Dependencies taxonomy) et pré-entraîné sur un grand corpus. Néanmoins, ce modèle ne donnait pas des résultats satisfaisants, notamment pour la partie étiquetage morpho-syntaxique (Part-of-Speech). Nous avons donc décidé d'entraîner notre propre modèle d'étiquetage morpho-syntaxique sur nos annonces immobilières.

A partir des dépendances syntaxiques, nous construisons le graphe de dépendances pour chaque phrase. Nous avons ensuite extrait le plus court chemin entre chaque paire d'entités candidates à une relation. Enfin, grâce à des règles prédéfinies, nous déterminons si les chemins extraits correspondent à une relation ou non.

2.3 Représentation des connaissances

La dernière étape de notre travail porte sur la manière de représenter et d'interroger la connaissance extraite. Nous avons choisi d'utiliser un graphe de connaissance car il offre une manière flexible de représenter les entités (nœuds) et les relations (arcs) mais aussi un langage de requête pour naviguer et raisonner sur les informations. Le modèle RDF et le langage de requête GeoSPARQL ont été choisis pour décrire et stocker les données.

3 Evaluation

Nous avons évalué le modèle d'extraction de connaissances à partir d'un jeu de données d'environ 1200 annonces immobilières préalablement traitées, nettoyées et découpées en 10 échantillons pour faire une validation croisée. Nous avons comparé plusieurs architectures de notre modèle avec le modèle Spacy pré-entraîné pour le français. Le meilleur modèle, qui utilise l'architecture *BiLSTM+CRF* avec l'embedding décrit dans 2.1, obtient un F1-Score de 0.876 soit 5.5 points au-dessus du modèle de Spacy pré-entraîné.

4 Conclusion et perspectives

Nous avons décrit dans cet article une méthode pour extraire des informations géospatiales des textes appliquée aux annonces immobilières écrites en français. Nous avons créé un modèle de reconnaissance d'entités pour extraire des lieux-nommés mais aussi les types de lieu, les entités spatio-temporelles et les modes de transport. Nous avons aussi conçu une méthode pour extraire des relations entre les entités et plus particulièrement des relations spatiales. Enfin, nous avons representé les connaissances extraites à l'aide d'un graphe de connaissance RDF.

Par la suite, nous envisageons de retrouver la localisation des lieux mentionnés et de les relier à des graphes de connaissances existants (e.g., GeoNamnes, DBpedia, etc.). Aussi, nous aimerions prendre en compte l'incertitude et l'imprécision des termes spatio-temporels afin d'améliorer la fiabilité de la localisation d'un lieu.

Références

[1] L. Cadorel et al., *Geospatial Knowledge in Housing Advertisements : Capturing and Extracting Spatial Information from Text.* In Proceedings of the 11th on Knowl- edge Capture Conference, K-CAP '21, page 41–48, New York, NY, USA, 2021. Association for Computing Machinery.

[2] B. Adams and K. Janowicz, *On the geo-indicativeness of non-georeferenced text.* In John G. Breslin, Nicole B. Ellison, James G. Shanahan, and Zeynep Tufekci, editors, Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012. The AAAI Press, 2012

[3] R. Grace, *Toponym usage in social media in emergencies.* International Journal of Disaster Risk Reduction, 52 :101923, 2021

[4] Y. Hu, et al., *A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements.* Int. J. Geogr. Inf. Sci., 33(4) :714–738, 2019.

Session 2 : Raffinement de graphes de connaissances - enrichissement et validation

Alignement entre sources : cas d'usage des plantes cultivées

F. Michel[3], F. Amardeilh[1], R. Bossy[2], C. Faron[3], C. Roussey[4], C. Noûs[5]

[1] Elzeard R&D, Pessac, France
[2] Université Paris-Saclay, INRAE, UR MAIAGE, 78350 Jouy-en-Josas, France
[3] Université Côte d'Azur, Inria, CNRS, I3S, 06902, Sophia-Antipolis, France
[4] Université Clermont Auvergne, INRAE, UR TSCF, Aubière, France
[5] Laboratoire Cogitamus, Havre de recherche, France.

florence.amardeilh@elzeard.co, robert.bossy@inrae.fr, faron@i3s.unice.fr, fmichel@i3s.unice.fr, catherine.roussey@inrae.fr, camille.nous@cogitamus.fr

Résumé

Dans cet article nous décrivons nos premiers travaux sur l'alignement de deux graphes de connaissances complémentaires utiles dans le domaine de l'agriculture : le thésaurus des usages des plantes cultivées (FCU) et le registre taxonomique national français TAXREF pour la faune, la flore et la fonge. Plusieurs méthodes d'alignement spécifiques à ce cas d'usage ont été implémentées. Les résultats montrent que dans ce domaine il sera nécessaire de nettoyer les alignements produits automatiquement.

Mots-clés

web sémantique, graphes de connaissances, alignement, taxonomie biologique, TAXREF-LD, thesaurus agricole, FrenchCropUsage, SKOS, plantes.

Abstract

In this article we describe our first work on the alignment of two complementary knowledge graphs useful in the agricultural domain. A SKOS thesaurus related to uses of cultivated plants (FCU), a knowledge graph of the biological taxonomy in France (TAXREF-LD). Several alignment methods specific to this use case have been implemented. The results show that for this use case it will be necessary to curate the alignments produced automatically.

Keywords

semantic web, knowledge graph, alignment, biological taxonomy, TAXREF-LD, agricultural thesaurus, FrenchCropUsage, SKOS, plant.

1 Introduction

Le projet ANR *Des Données aux Connaissances en Agronomie et Biodiversité* (D2KAB) illustre comment l'ingénierie des connaissances contribue au développement d'applications innovantes dans le domaine de l'agriculture. L'objectif de D2KAB est de créer un cadre pour transformer les données d'agronomie et de biodiversité en connaissances décrites sémantiquement, interopérables, exploitables et ouvertes. Pour construire un tel cadre, nous nous appuyons sur des ressources sémantiques (terminologies, vocabulaires, ontologies) pour décrire nos données et les publier en tant que données ouvertes liées [1]. Nous utilisons notamment le portail AgroPortal [8] pour trouver, publier et partager ces ressources sémantiques puis nous les exploitons dans des applications dédiées à l'agriculture ou l'environnement.

Alors que le web de données liées met à disposition de plus en plus de graphes de connaissances, leur réutilisation croisée reste souvent un défi. Cet article présente une méthode permettant d'aligner entre eux des graphes de connaissances représentant des points de vue différents sur les mêmes objets d'étude, afin de requêter conjointement ces graphes pour des raisons de complétude d'information. L'agriculture offre un cas d'usage particulier dans ce domaine, lié à la modélisation des plantes cultivées. Plusieurs expertises sont nécessaires pour décrire une plante cultivée : agriculteur versus agronome, agronome versus écologue. Le monde scientifique (les écologues ou agronomes) utilise les noms scientifiques issus de la science taxonomique pour désigner les organismes vivants (plantes, insectes). Ces noms scientifiques sont stockés dans des taxonomies biologiques. Le monde des usagers (les agriculteurs) utilise des noms vernaculaires pour désigner les organismes vivants qui interviennent dans leur pratique. De plus, une plante peut avoir plusieurs rôles en agriculture : (1) une plante cultivée dans un but de production, (2) une adventice (mauvaise herbe), (3) une plante cultivée dite plante de service, pour rendre un service à une autre plante cultivée, dite alors production principale (la plante de service est détruite sans être récoltée, au contraire de la plante de production).

Nous présentons tout d'abord les travaux sur l'alignement des taxonomies dans le domaine agricole. La section 3 décrit en détail des sources utilisées dans notre approche d'alignement. La section 4 présente les algorithmes d'alignements mis en oeuvre. La section 5 présente deux types d'évaluations effectuées sur les résultats de nos algorithmes. Enfin, nous concluons nos travaux en présentant des perspectives d'amélioration.

2 Travaux antérieurs

Il existe déjà plusieurs graphes de connaissances qui essaient de combiner les points de vue des agronomes et des agriculteurs. Le plus connu est le thésaurus Agrovoc de la FAO [2]. Nous débuterons par décrire les travaux sur l'alignement des taxonomies biologiques, qui ont été utilisées pour évaluer les outils d'alignement.

2.1 Les taxonomies biologiques

La taxonomie est la science de la diversité du vivant. Elle consiste à décrire les organismes vivants et à les organiser en groupes appelés *taxons*, selon une hiérarchie reflétant l'histoire de leur évolution [9]. Les taxons se situent à différents niveaux de généralité, appelés *rangs taxonomiques*, parmi lesquels on peut citer l'espèce, la famille, ou la variété.

Il existe de nombreux référentiels taxonomiques, ou taxonomies. La difficulté de leur maintenance par les curateurs, et *in fine* du choix d'un référentiel taxonomique tient, à la fois de leur volume et de la volatilité structurelle de leur contenu. En effet, à ce jour [3] recense près de deux millions d'espèces décrites, dont plus de 300.000 espèces de Magnoliophytes (*"plantes à fleurs"*) auxquelles appartiennent la majorité des plantes cultivées. De plus, le contour des unités conceptuelles des taxonomies est instable puisque les taxons et leur organisation constituent les hypothèses de travail des systématiciens. Cela signifie que les taxonomies sont soumises aux controverses scientifiques ; les taxonomies décrites à une date donnée peuvent être réfutées dans l'avenir.

Ce constat a aussi amené les systématiciens à réguler strictement les conventions de nommage des taxons. Les codes de nomenclature [14] permettent aux systématiciens de stabiliser la nomenclature face à la volatilité des taxonomies. Ces conventions génèrent différents types de recombinaisons pouvant induire des synonymes, voire des homonymes, faisant de la constitution, la maintenance et l'alignement des référentiels taxonomiques une tâche difficile.

2.1.1 Stratégies de curation

Nous mentionnerons trois référentiels parmi les plus complets et utilisés : NCBI Taxonomy, TAXREF, et Catalogue of Life. Ces trois référentiels couvrent bien les différents contours et les différentes politiques de curation.

NCBI Taxonomy [17] est la taxonomie de référence des bases de données maintenues par le NCBI, dont PubMed et GenBank. La taxonomie est complétée au gré des besoins selon les entrées des bases de données du NCBI. L'organisation hiérarchique est assurée par des curateurs volontaires qui se basent sur la littérature en systématique. La taxonomie du NCBI est constitutionnellement biaisée par l'abondance des études et reflète mal la biodiversité. De plus, une politique de stabilité des identifiants de taxons amène quelquefois des inexactitudes. Le principal avantage de NCBI Taxonomy reste les nombreux liens vers des bases de données moléculaires et bibliographiques.

TAXREF [7] est le référentiel taxonomique du Système d'Information de l'Inventaire National du Patrimoine naturel, diffusé et maintenu par le Muséum National d'Histoire Naturelle (MNHN). Il liste toutes les espèces recensées dans les territoires français (métropole et outre-mer), ainsi que plus de 650.000 noms scientifiques associés aux taxons de tous rangs taxonomiques. Les curateurs de TAXREF sont en contact direct avec les spécialistes identifiés pour chaque branche du vivant. TAXREF constitue donc une source primaire d'une taxonomie scientifiquement fondée pour les espèces que l'on trouve en France. TAXREF-LD est la distribution de TAXREF respectant les principes des données liées. Ce graphe de connaissance est décrit en détails dans la section 3.

Catalogue of Life [16] est un projet à l'ambition universelle, soutenu par le Global Biodiversity Information Facility (GBIF). Il s'agit d'une fédération de référentiels, chaque composant couvrant une branche du vivant (e.g. LPSN), un environnement (e.g. WoRMS) ou une zone géographique (e.g. ITIS). Cette stratégie permet d'obtenir un compromis entre exhaustivité et justesse.

2.1.2 Alignement entre taxonomies

L'alignement automatique des taxons issus de différents référentiels taxonomiques reste une question scientifique ouverte qui a notamment motivé plusieurs tâches dans la campagne Ontology Alignment Evaluation Initiative (OAEI) [1]. Les tâches *OAEI Taxon* et *Biodiv* portent sur la détection d'alignements entre taxons biologiques.

La tâche *OAEI Taxon* porte sur la détection d'alignements complexes. Le benchmark de *OAEI Taxon* est composé de quatre référentiels taxonomiques représentés sous forme de graphes de connaissances. En 2021, seulement 3 systèmes automatiques d'alignement sur 11 ont pu proposer des alignements valides : ATM, Fine-TOM et logmap [15]. De plus, la plupart des alignements proposés étaient jugés simples.

La tâche *Biodiv* porte sur la détection d'alignements simples dans le domaine de la biodiversité. L'un des benchmarks de *Biodiv* se compose de TAXREF-LD et NCBI Taxonomy. Malheureusement à cause de la taille de ces deux taxonomies, aucun système n'a été capable de proposer des alignements. A noter que l'outil AgreementMaker-Light (AML) a obtenu les meilleurs résultats sur les deux autres benchmarks de cette tâche [6].

En outre, les référentiels taxonomiques les plus connus et adoptés font rarement le travail de produire des alignements avec d'autres référentiels. Les auteurs de TAXREF ont fait ce travail en alignant TAXREF-LD et plusieurs autres référentiels dont NCBI Taxonomy. Ce calcul a été mis en oeuvre à l'aide de l'outil SILK [18] et d'une extension développée pour implémenter les règles métier d'alignement de noms scientifiques [2].

1. http://oaei.ontologymatching.org/
2. https://github.com/frmichel/taxrefmatch-silk-plugin

2.2 Sources associant des taxonomies biologiques et des noms vernaculaires de plantes cultivées

A notre connaissance, il existe trois sources qui proposent une représentation multiple des plantes cultivées avec une terminologie française : Agrovoc, la base de données mondiale EPPO et le catalogue du GEVES.

2.2.1 Agrovoc

Le thésaurus Agrovoc est publié par l'Organisation des Nations Unies pour l'Alimentation et l'Agriculture (FAO) [2]. Il est édité manuellement par une communauté mondiale d'experts et couvre tous les domaines d'intérêt de la FAO, y compris l'agriculture, la sylviculture, la pêche, l'alimentation et les domaines connexes. Il est disponible en 29 langues, avec une moyenne de 35 000 termes par langue et développé en SKOS-XL [12]. La force de ce thésaurus est sa couverture lexicale multilingue. Il est donc souvent utilisé pour annoter ou indexer des documents ou des images relatifs au domaine de l'agriculture. Agrovoc contient plusieurs représentations des plantes décrites dans des branches différentes de sa hiérarchie. Une plante peut être décrite par son nom scientifique. Par exemple, le `skos:Concept` http://aims.fao.org/aos/agrovoc/c_8283 représente l'espèce *"Vitis vinifera"*, qui a pour parent un `skos:Concept` représentant le genre *"Vitis"*. Une plante peut être décrite par son nom vernaculaire associé à la filière agricole qui la cultive. Par exemple le `skos:Concept` http://aims.fao.org/aos/agrovoc/c_3360 représente la vigne et a pour parent les cultures fruitières. Il existe plusieurs types de liens possibles entre ces deux branches d'Agrovoc, par exemple "includes" ou "is used as". La branche représentant les noms vernaculaires des plantes cultivées a plusieurs composants dont il n'est pas évident de comprendre la logique. Par exemple, le concept de vigne est associé :

- aux concepts de "vitis vinifera", "vitis labrusca", "vitis rotundifolia" et "vitis aestivalis" par un lien "includes",
- au concept de "vitis" par un lien "included in",
- au concept "beverage crop" par un lien "is used as",
- au concept de raisin par un lien "produces".

.

2.2.2 La base de données EPPO

La base de données mondiale EPPO [4] est maintenue par le Secrétariat de l'*Organisation Européenne et Méditerranéenne pour la Protection des Plantes* (EPPO) [3]. L'objectif de cette base est de fournir des informations spécifiques aux organismes nuisibles, qui ont été produites ou collectées par l'EPPO. Le contenu de la base de données est constamment mis à jour par le Secrétariat de l'EPPO. Cette base est interrogeable en ligne ou par le biais d'une API. Chaque plante est identifiée par un code de 5 caractères qui sert de référence dans de nombreuses autres bases de données agricoles européennes. Cette base identifie aussi des groupes de plantes (taxon biologique, filière agricole, ...) par des codes à 5 caractères. Par exemple le code pour l'espèce *"Vitis vinifera"* est VITVI. Le code pour le genre *"Vitis"* est 1VITG. EPPO représente aussi d'autres classifications des plantes (crop groups, commodity groups, crop destination,...). Les filières agricoles sont partiellement représentées dans la classification des crop groups (3CRGK). Cette classification contient par exemple les cultures fruitières (fruit crops : 3FRUC) et les légumes (vegetable crops : 3VEGC). Pour chacun de ces groupes sont affichés la liste des taxons associés. Ainsi l'espèce *"Vitis vinifera"* apparaît comme taxon associé des cultures fruitières. La filière vigne n'existe pas dans la classification crop groups.

2.2.3 Catalogue du GEVES

Le Groupe d'Etude et de contrôle des Variétés Et des Semences (GEVES) produit un catalogue officiel des espèces et variétés de plantes cultivées en France [4]. Ce catalogue contient 9 000 variétés pour 190 espèces. Toute variété, produite par un institut agricole est inscrite dans ce catalogue pour être commercialisé. Une variété de plante est décrite entre autre par son nom, le détenteur de la variété, une indication de son type variétal, son espèce biologique, sa filière agricole. Par exemple la variété de raisin de cuve abouriou, produite par Institut Français de la Vigne et du Vin, a comme indication de type variétal "couleur de baie blanche" et comme espèce *"Vitis vinifera"*. Ce catalogue est disponible sous forme de plusieurs fichiers CSV ou d'une API.

Ces trois sources montrent que le rang espèce des taxons biologiques est associé au nom vernaculaire de la plante pour identifier au mieux une plante cultivée. A notre connaissance cet article présente un premier cas d'étude d'alignement automatique de graphes de connaissances complémentaires dans le domaine de l'agriculture : un alignement de taxons biologiques avec des noms d'usage des plantes cultivées.

3 Sources à aligner : TAXREF-LD et FCU

Les deux graphes de connaissances que nous cherchons à aligner sont fondés sur le vocabulaire SKOS ou une extension de SKOS. TAXREF-LD est la publication du référentiel TAXREF sur le web de données liées. FCU est un thésaurus francophone des usages des plantes cultivées. Ces deux graphes complémentaires ont en commun uniquement les noms vernaculaires des plantes cultivées.

3.1 TAXREF et TAXREF-LD

TAXREF [7] est le référentiel taxonomique français pour la faune, la flore et la fonge. Outre un portail Web, un service REST et un ensemble de fichiers CSV téléchargeables, TAXREF est disponible sous forme d'un graphe de connaissances respectant les principes des données liées, nommé TAXREF-LD [11] [5].

3. https://www.eppo.int/

4. https://www.geves.fr/catalogue/

5. TAXREF-LD peut être téléchargé depuis https://doi.org/10.5281/zenodo.5876775. Il est possible de l'interroger par le biais d'un SPARQL endpoint public, et voir les informations disponibles dans le dépôt https://github.

Afin de refléter fidèlement la distinction entre taxonomie et nomenclature, TAXREF-LD comporte deux niveaux distincts de modélisation illustrés par la figure 1. Au niveau taxonomique, chaque taxon biologique est modélisé comme une classe OWL dont les membres sont les individus biologiques de ce taxon. La classe parente est le taxon de rang supérieur (e.g. *"Daucus carota"* est de rang espèce, la classe parente, *"Daucus"*, est de rang genre). Au niveau nomenclatural, les noms scientifiques sont représentés comme les concepts d'un thésaurus SKOS. Chaque nom (concept SKOS) est lié à un taxon (classe OWL) par une propriété indiquant s'il s'agit du nom de référence (nom *accepté* en zoologie ou *valide* en botanique), ou d'un synonyme.

Outre l'information strictement taxonomique, TAXREF-LD représente également d'autres types d'information : noms vernaculaires, habitats, statuts de conservation, statuts biogéographiques, interactions entre espèces, ainsi que les références bibliographiques associées à ces différentes informations. A noter que TAXREF-LD associe parfois le même nom vernaculaire à plusieurs taxons. Ces noms vernaculaires sont issus des publications où sont déclarés les noms scientifiques. Par ailleurs, TAXREF-LD est lié à plusieurs référentiels taxonomiques tierces dont Agrovoc et NCBI Organismal Taxonomy.

3.2 Thésaurus agricole FCU

Le thésaurus intitulé "usages des plantes cultivées en France" ou French Crop Usage (FCU) [6] normalise les noms de plantes cultivées en français. De plus, il les organise dans des catégories représentant des filières agricoles : par exemple, *"fourrage"* et *"grandes cultures"* sont deux exemples de filières agricoles. Ainsi, une hiérarchie est formée par des relations de généralisation/spécialisation entre les filières agricoles et les noms d'usage des plantes cultivées : par exemple, *"grandes cultures"* se spécialise en *"céréales"*). Les termes du thésaurus ont été sélectionnés manuellement à partir de documents de référence. Les documents étudiés pour construire le thésaurus sont :

Les statistiques agricoles annuelles de l'Agreste [7], les métadonnées du registre parcellaire graphique, le classement des plantes cultivées par groupe d'usage proposé par wikipédia France, le catalogue officiel des espèces et variétés de plantes cultivées en France du GEVES, les fiches "les plantes fourragères pour les prairies" du GNIS [8], la base Ephy qui décrit l'usage des produits phytosanitaires sur les plantes, le Larousse Agricole.

Concernant les légumes et leur classification, les points communs entre plusieurs sources ont été cherchés : Wikipedia, Bonduelle, FranceAgriMer, Encyclopedia Universalis, La ferme du Bec Hellouin. Notons qu'il n'existe pas de consensus sur la classification des légumes. Le choix des noms d'usage des plantes cultivées, les définitions associées et leur organisation ont été discutés par au moins un expert de la filière agricole. Ce thésaurus n'est pas complet et évolue en fonction des projets.

Le thésaurus est modélisé à l'aide du vocabulaire SKOS proposé par le W3C [13], la figure 2 en présente un extrait. Il est disponible sur le web de données liées [9]. FCU contient 526 `skos:Concepts`. La profondeur maximale de la hiérarchie est de 6. Chaque concept est défini par un ensemble d'étiquettes (les noms vernaculaires de la plante), des notes, des liens vers d'autres sources d'information et de liens hiérarchiques. Nous présentons une liste succincte des propriétés utilisées pour définir chaque `skos:Concept` :

- `skos:prefLabel` : contient le terme utilisé comme étiquette préférée du concept en français. En général, le terme est le nom vernaculaire de la plante cultivée avec une indication de son usage (ex : *"vigne ornementale"*, *"vigne cultivée"* ou *"vigne de cuve"*). Au besoin, cette étiquette préférée peut être construite de manière artificielle pour bien identifier l'usage agricole de la plante. En générale, cette étiquette est prise dans l'une des sources identifiées. Par exemple, les trois étiquettes (*"vigne ornementale"*, *"vigne cultivée"* ou *"vigne de cuve"*) ont toutes été trouvées dans une des sources.
- `skos:altLabel` : contient les autres termes qui peuvent être utilisés comme étiquettes du concept (ex : *"vigne vierge"* est une autre étiquette de *"vigne ornementale"*). Ces étiquettes peuvent indiquer le produit récolté (ex : "raisin") ou l'activité agricole (ex : viticulture).
- `skos:definition` : contient la définition en français du concept justifiant sa position dans la hiérarchie du thésaurus.
- `rdfs:seeAlso` : contient un lien web vers une définition retenue lors de la construction du thésaurus, comme par exemple les pages wikipédia.
- `skos:note` contient au moins une définition trouvée dans une autre source comme l'Agreste ou wikipédia.

Lorsqu'une plante a plusieurs usages, elle est représentée par plusieurs concepts : un concept pour chacun des usages, plus un concept pour l'ensemble des usages, parent des concepts précédents. Un concept dans la branche "multiusage" porte le nom vernaculaire de la plante sans indication d'usage (par exemple *"carotte"*). Ce concept est ensuite décliné en autant de fils qu'il y a d'usages (*"carotte potagère"* pour l'alimentation humaine et *"carotte fourragère"* pour l'alimentation animale). Chacun des fils est de plus positionné à un seul endroit dans la branche *"usage des plantes cultivées"*. Dans la figure 3 le concept *"carotte potagère"* est positionné comme fils du concept *"légume racine"*.

3.3 Difficultés pour aligner les sources

Bien que les taxonomies et listes de plantes cultivées référencent les mêmes objets du monde (les organismes vivants), leur alignement présente plusieurs difficultés qui rendent nécessaire une validation manuelle en pratique.

La taxonomie est une science difficile à appréhender pour

com/frmichel/taxref-ld/.

6. https://doi.org/10.15454/QHFTMX
7. l'Agreste est le service statistique ministériel de l'agriculture
8. Le GNIS est l'interprofession des semences et plants, il a été renomé SEMAE
9. http://ontology.irstea.fr/pmwiki.php/Site/FrenchCropUsage

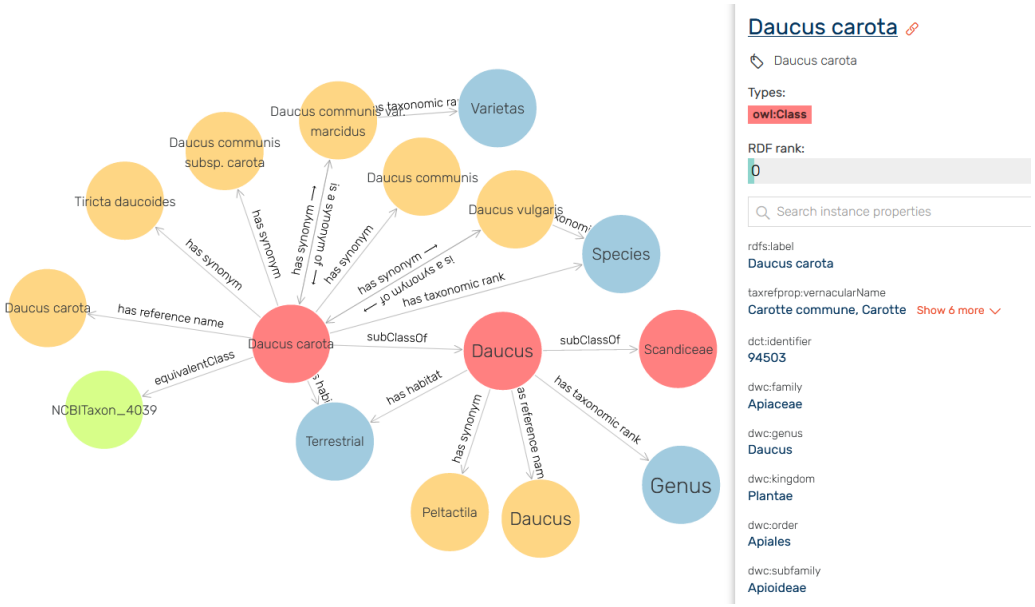


FIGURE 1 – Deux niveaux de modélisation dans TAXREF-LD, exemple de la carotte. Noeuds rouges : taxons modélisés comme des classes OWL. Noeuds oranges : noms scientifiques modélisés comme des concepts SKOS associés aux taxons en tant que nom de référence ou synonyme.

les non-spécialistes. Un taxon est une hypothèse scientifique affirmant qu'un ensemble d'individus biologiques appartiennent au même groupe taxonomique en raison de certaines caractéristiques communes. Dans le cas simple, à un taxon est associé un nom scientifique. Cependant, l'évolution du consensus scientifique entraîne des changements dans la taxonomie, ainsi des recombinaisons peuvent se produire : deux taxons peuvent être fusionnés en un seul, un taxon existant peut être divisé en deux taxons distincts, ou un taxon peut changer de rang taxonomique (espèce vers sous-espèce par exemple). Par conséquent, un taxon peut avoir un nom préféré utilisé pour désigner le taxon, et plusieurs synonymes. Les noms et leurs recombinaisons sont publiés dans la littérature scientifique, toutefois la prise en compte de ces évolutions dans les taxonomies et les liste de plantes cultivées peut se faire à des rythmes différents, ce qui mène fréquemment à des désaccords. Par exemple, une liste de plantes peut utiliser un nom scientifique qui n'est plus le nom de référence du taxon, ou dont le taxon a changé de taxon parent ou de rang.

Les Codes de nomenclature regroupent l'ensemble des règles régissant les noms scientifiques. Si ces règles s'appliquent sans ambiguïté jusqu'aux niveaux espèce et sous-espèce, les noms scientifiques associés aux rangs inférieurs (e.g. variété, cultivar) ne sont pas concernés. Or les listes de plantes cultivées dénotent parfois des taxons appartenant à ces niveaux inférieurs. En outre, il n'existe pas de règle sur le fait qu'un nom de plante cultivée dénote une sous-espèce, une variété, etc., et les noms vernaculaires retenus pour nommer les plantes sont parfois spécifiques à une région donnée, rendant l'alignement encore plus délicat. A ces difficultés s'ajoutent des problèmes de fiabilité. En raison de leur complexité, les règles de nomenclature ne sont pas toujours respectées. Par exemple le catalogue du GEVES donne l'autorité sans la date (e.g. "L." au lieu de "L. 1758") et ne respecte pas la casse. Qui plus est, les listes de plantes cultivées sont construites par agrégation mais ne précisent pas nécessairement leurs sources (publications scientifiques attestant de l'utilisation d'un nom), empêchant d'évaluer la confiance que l'on peut leur accorder.

Enfin, il existe des difficultés plus techniques, liées aux choix de modélisation des taxons, noms scientifiques et cultures. Par exemple, TAXREF-LD sépare strictement taxonomie et nomenclature. D'autres classifications ne font pas cette distinction, représentant à la fois des taxons et leurs noms. Certaines classifications ne représentent que des noms scientifiques, comme Catalog of Life. Les listes de plantes cultivées ne retiennent souvent qu'un nom scientifique en lieu et place d'un taxon, nom qui n'est peut-être plus le nom de référence du taxon. Ces variations de conception et de modélisation posent ainsi des questions récurrentes quant au choix des objets à aligner : aligne-t-on deux taxons, un taxon et un nom, une plante cultivée et un taxon etc. ?

4 Algorithme d'alignement mis en oeuvre

La section 3.3 a souligné les différences de modélisation existant entre les taxonomies biologiques et les listes de plantes cultivées, ainsi que l'écart entre les données représentées (noms vernaculaires, nom scientifique, taxon). Ces différences rendent difficile l'utilisation d'outils classiques d'alignement d'ontologies ou de liage d'entités. Aussi nous avons exploré puis combiné plusieurs méthodes.

Le code implémentant les méthodes décrites dans cette sec-

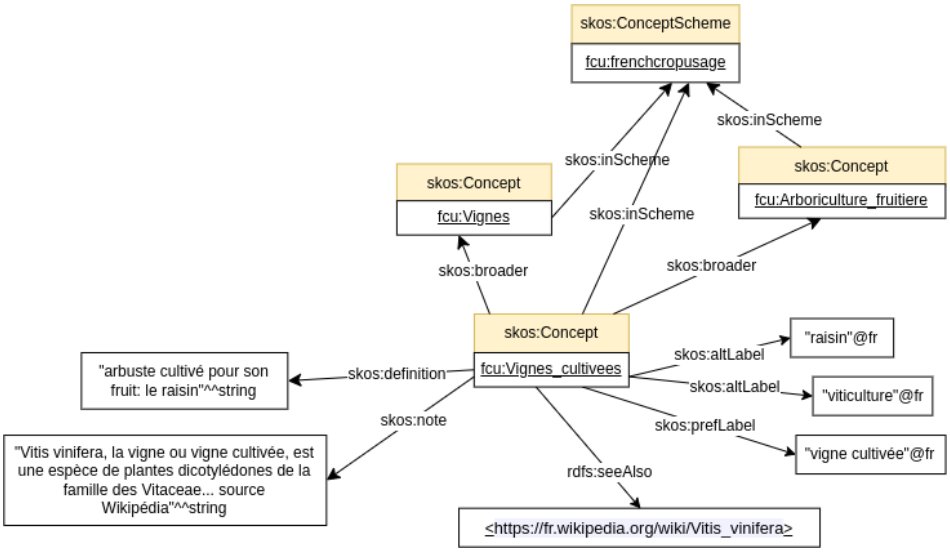


FIGURE 2 – Un extrait du thésaurus FCU présentant le concept de vigne cultivée.

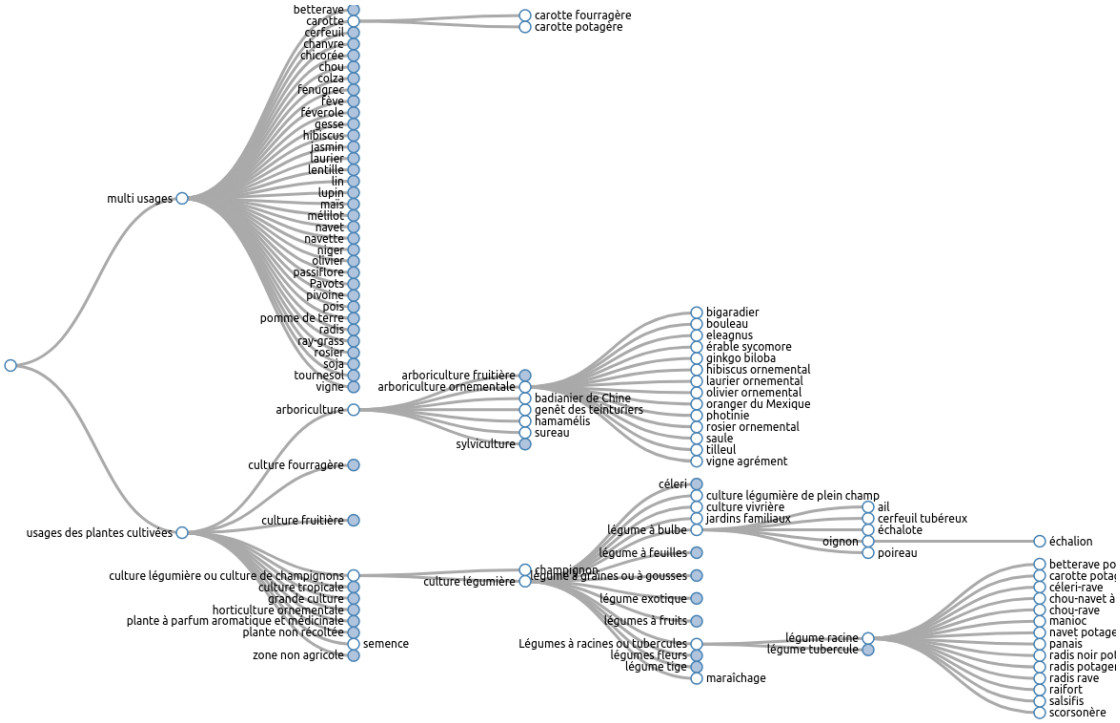


FIGURE 3 – Un exemple de visualisation du thésaurus avec l'outil SKOSPlay

tion ainsi que les données produites sont disponibles sous licence ouverte sur un dépôt public [10].

4.1 Méthodes d'alignement

Une première méthode consiste en un alignement direct entre FCU et TAXREF-LD basé sur la correspondance exacte, insensible à la casse, entre les noms d'usage des plantes cultivées de FCU (étiquettes préférées ou alternatives de concepts) et les noms vernaculaires de TAXREF-LD. Pour ce faire nous avons utilisé la version 2.2 de FCU et la version 15 de TAXREF-LD. En pratique, cette méthode donne des résultats médiocres en raison de la variabilité des noms vernaculaires retenus dans chaque source.

Une autre approche consiste à utiliser une source intermédiaire faisant la correspondance entre les noms vernaculaires des plantes cultivées et leurs noms scientifiques. La section 3 propose plusieurs référentiels faisant autorité, qui font l'objet d'une curation manuelle. Nous en avons retenu deux : la *Base de Données Mondiale* publiée par l'EPPO [4] que nous avons interrogée par son interface programmatique ; le *Catalogue officiel des espèces et variétés de plantes cultivées en France* publié par le GEVES [11] dont nous avons téléchargé les fichiers tabulaires depuis le site web du GEVES. Dans un premier temps, l'algorithme cherche à faire correspondre les noms d'usage des plantes de FCU avec les noms vernaculaires de ces deux référentiels. Dans le cas du GEVES, il s'agit d'une correspondance exacte, insensible à la casse. Dans le cas d'EPPO, il s'agit d'une correspondance approchée implémentée par l'API EPPO [12], toutefois la mesure de distance utilisée n'est pas documentée. L'algorithme retient le nom scientifique correspondant à chaque nom vernaculaire. Dans un deuxième temps, il cherche ce nom scientifique dans TAXREF-LD, puis il retient le taxon dont ce nom scientifique est soit le nom de référence soit un synonyme. La base EPPO et TAXREF-LD respectent strictement le code de nomenclature pour le nommage des noms scientifiques (sous la forme : "nom latin autorité, année", e.g. *"Prunus armeniaca L., 1753"*. La correspondance est donc une simple égalité insensible à la casse. En revanche, le catalogue du GEVES ne fournit pas l'année et ne respecte pas la casse (e.g. *"prunus armeniaca l."*). La correspondance se fait donc en cherchant des noms scientifiques de TAXREF-LD commençant par le nom issus du GEVES du GEVES (comparaison insensible à la casse).

Afin de permettre à des experts de valider les alignements, les résultats des trois méthodes sont conservés (alignement direct, via EPPO, via GEVES) et ordonnés par étiquette de FCU. Un score de confiance est attribué à chaque alignement candidat, calculé en fonction du nombre de méthodes (1, 2 ou 3) ayant proposé cet alignement. Le score peut donc valoir 1/3, 2/3 ou 1. Notons que le score 1 signifie simplement l'accord entre les trois méthodes, mais ne garantit pas sa justesse qui doit être vérifiée par un expert.

10. https://github.com/Wimmics/d2kab-alignments
11. https://www.geves.fr/catalogue/
12. Service /tools/names2codes : https://data.eppo.int/documentation/rest

4.2 Choix des entités à aligner

Dans les trois méthodes ci-dessus, on cherche à aligner les concepts de FCU avec des taxons de TAXREF-LD. Côté TAXREF-LD, on restreint les taxons candidats aux rangs espèce ou infra-spécifiques (sous-espèce, variété, etc.). En effet, le nom scientifique d'une plante cultivée se caractérise au moins par son espèce.

Côté FCU, on considère deux groupes de concepts de FCU à aligner. Dans le groupe *plantes cultivées*, on ne considère que les plantes de la branche "usages des plantes cultivées" et seulement celles des deux derniers niveaux de la hiérarchie (les feuilles ou leurs parents immédiats). Nous faisons donc l'hypothèse que l'unité d'alignement avec TAXREF-LD est un usage précis de plantes cultivées et non un regroupement de plantes (comme céréales). Dans le groupe *tous concepts*, on considère tous les concepts des branches "usages des plantes cultivées" et "multiusage" quel que soit leur niveau dans la hiérarchie. On ne fait donc aucune hypothèse sur l'unité d'alignement entre FCU et TAXREF-LD. Le groupe *plantes cultivées* est donc un sous-ensemble du groupe *tous concepts*.

5 Évaluation des alignements

Les résultats des méthodes d'alignement ont été évalués de deux manières.

5.1 Évaluation quantitative

Les statistiques données dans cette section ont été calculées par des requêtes SPARQL soumises depuis deux Jupyter Notebooks disponibles sur le dépôt du projet [13].

L'algorithme d'alignement a été exécuté pour les deux groupes de concepts décrits en section 4.2. Dans le groupe *plantes cultivées* qui contient 447 concepts, l'algorithme a proposé 651 alignements pour 300 de ces concepts (67% des concepts alignés) vers 579 taxons. Aucun alignement n'a été proposé pour 147 concepts (33%). Ces 147 concepts non alignés ont été évalués par un expert qui a indiqué que 118 concepts auraient dû être alignés car ils correspondent bien à des plantes cultivées et non à des groupes de plantes. Dans le groupe *tous concepts* qui considère 526 concepts, l'algorithme a proposé 710 alignements pour 337 concepts (64% des concepts alignés) vers 609 taxons. Aucun alignement n'a été proposé pour 189 concepts (36% des concepts non alignés). Le détail des nombres d'alignements proposés par méthode et par groupe est donné dans la table 1.

On remarque que la méthode d'alignement direct fournit de nombreux alignements mais est peu discriminante : elle génère en moyenne 2,39 alignements/concept pour 86% des concepts dans le groupe *plantes cultivées*, et 2,33 alignement/concept pour 77% des concepts dans le groupe *tous concepts*. A l'inverse, la méthode utilisant le catalogue du GEVES est plus discriminante - environ 1 alignement/concept - mais pour seulement 14% et 15% des concepts respectivement. La méthode utilisant EPPO semble la plus équilibrée : en moyenne 1,36 alignements/concept pour

13. Notebooks query-alignments.ipynb disponibles sur https://github.com/Wimmics/d2kab-alignments

TABLE 1 – Nombre d'alignements produits, et nombre de concepts et taxons impliqués dans ces alignements. La ligne "Total dédoubl." donne le total dans chaque groupe après suppression des alignements proposés par plusieurs méthodes.

Méthode	Nb. total d'alignements	Nb. de concepts FCU	Nb. de taxons
Groupe *plantes cultivées*			
Align. direct	385	161	369
via cat. GEVES	67	64	57
via BD EPPO	362	266	315
Total dédoubl.	**651**	**300**	**579**
Groupe *tous concepts*			
Align. direct	406	174	385
via cat. GEVES	84	81	70
via BD EPPO	404	300	336
Total dédoubl.	**710**	**337**	**609**

TABLE 2 – Nombre d'alignements proposés par 2 ou 3 méthodes à la fois.

Communs aux 3 méthodes	direct & GEVES	direct & EPPO	GEVES & EPPO
Groupe *plantes cultivées*			
15	17	115	46
Groupe *tous concepts*			
18	20	123	59

67% des concepts du groupe *plantes cultivées*, et 1,34 alignement/concept pour 64% des concepts du groupe *tous concepts*.

En outre, une analyse détaillée indique que les trois méthodes sont fortement complémentaires. En effet, quel que soit le groupe, environ 77% des alignements ne sont proposés que par une seule méthode (503 alignements dans le groupe *plantes cultivées*, 544 dans le groupe *tous concepts*). La table 2 montre que les trois méthodes ne s'accordent que sur 15 alignements (2,3%) dans le groupe *plantes cultivées*, et 18 alignements (2,5%) dans le groupe *tous concepts*. L'accord le plus fort apparaît entre les méthodes d'alignement direct et via EPPO avec seulement 115 alignements (17%), 123 alignements (17%) respectivement.

5.2 Évaluation qualitative

L'évaluation qualitative porte sur un sous-ensemble de plantes : la vigne, la carotte, les salades, la tomate. Pour chaque plante, deux experts ont évalué les couples (concept FCU, taxon TAXREF-LD) existants et détecté les couples manquants. Sur ce faible nombre d'alignements, les experts étaient majoritairement d'accord. Aucune avis contradictoire entre experts n'a été noté. La différence vient de couples manquants proposé par un des experts. Les alignements ont été qualifiés à l'aide de propriétés skos match. Ce choix est uniquement pragmatique et nous avons précisé la signification de ces propriétés de la manière suivante :

— `skos:exactMatch` signifie dans notre cas que le groupe de plantes représenté par le taxon est utilisé pour remplir cet usage en agriculture. Par exemple, la sous-espèce *"Vitis vinifera subsp. vinifera"* a pour usage "vigne cultivée".
— `skos:broadMatch` signifie que le groupe de plantes représenté par l'usage est inclus dans le groupe de plantes représenté par le taxon. Par exemple, les plantes qui ont pour usage "vigne cultivée" sont toutes de l'espèce *"Vitis vinifera"*.
— `skos:narrowMatch` signifie que le groupe de plantes représenté par l'usage inclut l'ensemble des plantes représenté par le taxon. Par exemple, les chicorées potagères incluent la variété *"Cichorium intybus var. sativum"*.
— `skos:closeMatch` signifie qu'il existe un lien entre le groupe de plantes représentées par l'usage et celui du taxon mais que la signification de ce lien est inconnue de l'expert.

Dans un futur proche nous définirons un ensemble de propriétés spécifiques à l'alignement entre un taxon biologique et un usage de plante en agriculture.

Le tableau 3 présente l'évaluation des alignements de l'ensemble des méthodes sur le groupe *plantes cultivées*. Pour le sous-ensemble de plantes considéré pour l'évaluation, la base EPPO a produit 15 alignements (dont 2 faux), et le catalogue du GEVES a produit 2 alignements. 9 alignements ont été trouvés en direct (dont 2 faux). 7 alignements sont communs à EPPO et en direct (dont 1 faux). 2 alignements sont communs au catalogue du GEVES et à EPPO.

Le tableau 4 présente les évaluations des alignements de l'ensemble des méthodes sur le groupe *tous concepts*. Sur ce groupe, la base EPPO a produit 17 alignements (dont 2 faux), et le catalogue du GEVES a produit 3 alignements. 10 alignements ont été trouvés en direct (dont 2 faux). 7 alignements sont communs à EPPO et en direct (dont 1 faux). 3 alignements sont communs au catalogue du GEVES et à EPPO.

Dans le cas de la vigne, il existe plusieurs alignements de type exact match entre un même concept FCU et plusieurs taxons. Cela signifie qu'une plante cultivée correspond à plusieurs espèces ou que certains taxons sont référencés plusieurs fois par des noms synonymes. Dans le cas des salades nous avons noté l'inverse, il existe plusieurs alignements de type exact match entre un même taxon et plusieurs concepts FCU. Cela signifie que la même espèce est utilisée pour différents usages.

EPPO est la source qui produit le plus d'alignements mais certains d'entre eux sont jugés erronés par les experts. Le catalogue du GEVES produit peu d'alignements mais ils sont tous justes. L'accord entre deux sources n'est pas un bon critère pour nettoyer les alignements étant donné qu'un des alignements jugés faux a été détecté par EPPO et en direct. Cette évaluation qualitative n'a pas identifié d'accord entre les 3 sources. Nous avons besoin de procéder à plus d'évaluation pour identifier si le catalogue du GEVES est bien la source de référence à utiliser. Nous aurons aussi besoin d'étudier pourquoi cette source, qui recense toutes les

TABLE 3 – Résultat de l'évaluation qualitative pour le groupe *plantes cultivées*

nom de cultures	nb alig. détectés	nb alig. exact	nb alig. broad	nb alig. narrow	nb alig. close	nb alig. faux	nb alig. manquants
salade	22	11	5	2	0	4	0
tomate	2	2	0	0	0	0	0
carotte	0	0	0	0	0	0	2
vigne	3	2	1	0	0	0	4

TABLE 4 – Résultat de l'évaluation qualitative pour le groupe *tous concepts*

nom de cultures	nb alig. détectés	nb alig. exact	nb alig. broad	nb alig. narrow	nb alig. close	nb alig. faux	nb alig. manquants
salade	24	11	6	2	0	5	0
tomate	2	2	0	0	0	0	0
carotte	3	1	2	0	0	0	0
vigne	3	2	1	0	0	0	4

variétés cultivées, produit si peu d'alignements.

6 Conclusion et Perspectives

Les méthodes d'alignements automatiques que nous avons produites ne donnent pas entièrement satisfaction. Plusieurs causes peuvent être identifiées : la variabilité des noms vernaculaires qui ne suivent aucune convention, le manque de couverture en noms vernaculaires des taxonomies biologiques, et la simplicité des techniques de comparaison mises en oeuvre actuellement dans notre algorithme. Concernant ce dernier point, nous envisageons d'améliorer l'algorithme en utilisant des mesures de similarité plus adaptées. Par exemple, en utilisant une distance de Levenstein pour la correspondance entre noms vernaculaires, ou les règles métier d'alignement de noms scientifiques implémentées pour aligner TAXREF-LD avec d'autres référentiels taxonomiques (voir section 2.1.2). Ainsi, les alignements produits ont besoin d'être améliorés et nettoyés par des experts. Le catalogue du GEVES est la source qui a produit les alignements les plus fiables (validés par les experts) mais en nombre insuffisant.

Nos travaux montrent que les alignements entre des classifications agricoles et des taxonomies biologiques sont plus complexes que de simples correspondances 1:1. Nous avons besoin d'exprimer le fait qu'une plante cultivée correspond à plusieurs espèces, voire à un ensemble d'espèces et de sous-espèces, et inversement. Il s'agit donc d'alignements N:N pouvant impliquer différents types de relations.

Nous avons étudié à ce jour deux schémas permettant de stocker les alignements : le langage EDOAL et le schéma "A Simple Standard for Sharing Ontology Mappings" (SSSOM). EDOAL permet de représenter les alignements complexes [5] mais est difficile d'accès pour les non-spécialistes de ce langage. Il faudra donc réfléchir à des modalités de validation des alignements pour les agronomes. SSSOM est un standard en cours d'évolution qui pour le moment se limite aux alignements simples 1:1 [10]. La prochaine étape de notre travail est de définir un schéma permettant d'exprimer nos alignements automatiques et leurs validations par des experts.

Références

[1] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *Semantic Web and Information Systems*, 5(3) :1–22, 2009.

[2] Caterina Caracciolo, Armando Stellato, Ahsan Morshed, Gudrun Johannsen, Sachit Rajbhandari, Yves Jaques, and Johannes Keizer. The AGROVOC linked dataset. *Semantic Web - Interoperability, Usability, Applicability*, 4(3) :341–348, 2013. http://content.iospress.com/articles/semantic-web/sw106.

[3] Arthur D Chapman. Numbers of living species in Australia and the world. https://www.awe.gov.au/science-research/abrs/publications/other/numbers-living-species, Canberra, Australia, september 2009.

[4] EPPO. EPPO Global Database (available online). https://gd.eppo.int, 2022.

[5] Jérôme Euzenat, François Scharffe, and Antoine Zimmermann. Expressive alignment language and implementation. Contract, June 2007. https://hal.inria.fr/hal-00822892.

[6] Daniel Faria, Beatriz Lima, Marta Contreiras Silva, Francisco M Couto, and Catia Pesquita. AML and AMLC results for OAEI 2021. In *Proceedings of the 16th International Workshop on Ontology Matching co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual conference*, pages 131–136, 2021.

[7] Olivier Gargominy, Sandrine Tercerie, C Régnier, T Ramage, P Dupont, P Daszkiewicz, and L Pon-

cet. TAXREF v15, référentiel taxonomique pour la France : méthodologie, mise en œuvre et diffusion. https://inpn.mnhn.fr/programme/referentiel-taxonomique-taxref, 2021.

[8] Clement Jonquet, Anne Toulet, Elizabeth Arnaud, Sophie Aubin, Esther Dzalé Yeumo, Vincent Emonet, John Graybeal, Marie-Angélique Laporte, Mark A. Musen, Valeria Pesce, and Pierre Larmande. AgroPortal : a vocabulary and ontology repository for agronomy. *Computers and Electronics in Agriculture*, 144 :126–143, January 2018.

[9] Guillaume Lecointre and Hervé Le Guyader. *Classification phylogénétique du vivant : tome 2*. Belin, 05 2017.

[10] Nicolas Matentzoglu, James P. Balhoff, Susan M. Bello, Chris Bizon, Matthew Brush, Tiffany J. Callahan, Christopher G Chute, William D. Duncan, Chris T. Evelo, Davera Gabriel, John Graybeal, Alasdair Gray, Benjamin M. Gyori, Melissa Haendel, Henriette Harmse, Nomi L. Harris, Ian Harrow, Harshad Hegde, Amelia L. Hoyt, Charles T. Hoyt, Dazhi Jiao, Ernesto Jiménez-Ruiz, Simon Jupp, Hyeongsik Kim, Sebastian Koehler, Thomas Liener, Qinqin Long, James Malone, James A. McLaughlin, Julie A. McMurry, Sierra Moxon, Monica C. Munoz-Torres, David Osumi-Sutherland, James A. Overton, Bjoern Peters, Tim Putman, Núria Queralt-Rosinach, Kent Shefchek, Harold Solbrig, Anne Thessen, Tania Tudorache, Nicole Vasilevsky, Alex H. Wagner, and Christopher J. Mungall. A Simple Standard for Sharing Ontological Mappings (SSSOM). dec 2021. http://arxiv.org/abs/2112.07051.

[11] Franck Michel, Olivier Gargominy, Sandrine Tercerie, and Catherine Faron-Zucker. A Model to Represent Nomenclatural and Taxonomic Information as Linked Data. Application to the French Taxonomic Register, TAXREF. In *Proceedings of the ISWC2017 workshop on Semantics for Biodiversity (S4BioDiv)*, volume 1933, Vienna, Austria, 2017. CEUR Workshop Proceedings. http://ceur-ws.org/Vol-1933/paper-3.pdf.

[12] Alistair Miles and Sean Bechhofer. SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL). https://www.w3.org/TR/skos-reference/skos-xl.html, 2009.

[13] Alistair Miles and Sean Bechhofer. SKOS simple knowledge organization system reference. http://www.w3.org/TR/skos-reference/, 2009.

[14] Dan H. Nicolson. A history of botanical nomenclature. *Annals of the Missouri Botanical Garden*, 78(1) :33–56, 1991. http://www.jstor.org/stable/2399589.

[15] Mina Abd Nikooie Pour, Alsayed Algergawy, Florence Amardeilh, Reihaneh Amini, Omaima Fallatah, Daniel Faria, Irini Fundulaki, Ian Harrow, Sven Hertling, Pascal Hitzler, Martin Huschka, Liliana Ibanescu, Ernesto Jiménez-Ruiz, Naouel Karam, Amir Laadhar, Patrick Lambrix, Huanyu Li, Ying Li, Franck Michel, Engy Nasr, Heiko Paulheim, Catia Pesquita, Jan Portisch, Catherine Roussey, Tzanina Saveta, Pavel Shvaiko, Andrea Splendiani, Cássia Trojahn, Jana Vatascinová, Beyza Yaman, Ondrej Zamazal, and Lu Zhou. Results of the Ontology Alignment Evaluation Initiative 2021. In Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, and Cássia Trojahn, editors, *Proceedings of the 16th International Workshop on Ontology Matching co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual conference, October 25, 2021*, volume 3063 of *CEUR Workshop Proceedings*, pages 62–108. CEUR Workshop Proceedings, 2021. http://ceur-ws.org/Vol-3063/oaei21_paper0.pdf.

[16] Y. Roskov, G. Ower, T. Orrell, D. Nicolson, N. Bailly, and et al. (eds) Kirk, P. M. Species 2000 & ITIS Catalogue of Life, 2019 Annual Checklist. http://www.catalogueoflife.org/annual-checklist/2019/, 2019.

[17] Conrad L Schoch, Stacy Ciufo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard Mcveigh, Kathleen O'Neill, Barbara Robbertse, et al. NCBI Taxonomy : a comprehensive update on curation, resources and tools. *Database : the journal of biological databases and curation*, 2020 :21, 2020. https://doi.org/10.1093/database/baaa062.

[18] Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. Silk - A Link Discovery Framework for the Web of Data. In *2nd Workshop about Linked Data on the Web*, volume 538, Madrid, Spain, 2009. CEUR Workshop Proceedings. http://ceur-ws.org/Vol-538/ldow2009_paper13.pdf.

Remerciements

Ce travail a été réalisé avec le soutien du projet "Des Données aux Connaissances en Agronomie et Biodiversité (D2KAB–www.d2kab.org) financé par l'Agence Nationale de la Recherche (ANR-18-CE23-0017) et du projet "Partage de Connaissances" (PACON) du programme transverse MetaBio financé par INRAE; ainsi qu'avec l'aide de l'entreprise Elzeard https://www.elzeard.co. Nous remercions également les membres de la tâche 4.3 du projet D2KAB : Sophie Aubin, Stephan Bernard, Sonia Bravo, Anna Chepaikina, Baptiste Darnala, Matthieu Hirschy, Clement Jonquet et Nadia Yacoubi. Un remerciement particulier pour Juliette Raphel, ingénieur agronome chez Elzeard et les deux agronomes spécialistes de la vigne qui ont participé à l'évaluation : Thierry Lacombe de SupAgro et Olivier Yobregat de l'Institut Français de la Vigne et du Vin (IFV).

IOPE: Interactive Ontology Population and Enrichment Guided by Ontological Constraints

S. Baghernezhad-Tabasi[1] L. Druette[2] F. Jouanot[1] C. Meurger[2] MC. Rousset[1,3]

[1] Université Grenoble Alpes CNRS, LIG
[2] Université Claude Bernard Lyon 1, SAMSEI
[3] Institut Universitaire de France Paris

Mots-clés

Ontology Engineering, Knowledge Acquisition, Automation Form Generation, Simulation-based Training in Medecine.

1 Introduction

Dans cet article, nous abordons la construction d'ontologies métiers pour capturer les compétences de spécialistes experts du domaine, et cela dans un but de partage avec une communauté d'apprenants et de spécialistes moins expérimentés. Les ontologies sont l'épine dorsale de nombreux systèmes d'information qui nécessitent un accès à des connaissances structurées. De par leur nature même, les ontologies du monde réel sont des artefacts dynamiques qui évoluent à la fois dans leur structure (le modèle de données) et leur contenu (les instances). Les maintenir à jour est une opération critique pour la plupart des applications qui s'appuient sur les technologies du Web sémantique. Ces mises à jour englobent à la fois les aspects d'enrichissement et de peuplement des ontologies. L'enrichissement consiste à étendre une ontologie avec de nouveaux concepts et de nouvelles relations sémantiques, tandis que le peuplement consiste à ajouter de nouvelles instances. Construite sur la base de documentation métier, la mise à jour d'une ontologie est généralement effectuée de manière exploratoire et requière des interactions nombreuses avec l'expert pour prendre en compte les connaissances non documentées. Cependant, ces mises à jour manuelles pèsent sur les experts et rendent l'ensemble de l'écosystème ontologique inefficace. Dans cet article, nous préconisons une approche alternative et plus efficace, et proposons de gérer automatiquement les mises à jour grâce à des interactions avec l'expert via une interface utilisateur.

Cette approche de mises à jour automatiques et interactives présentent deux défis : **1)** Alors que les ontologies sont généralement représentées sous la forme de graphes, il est intrinsèquement difficile et contre-intuitif de fournir une représentation graphique des ontologies à l'usage des experts. **2)** Il est difficile d'évaluer comment les experts pourraient réaliser des mises à jour sur une ontologie d'une manière interactive, sans qu'ils possèdent une connaissance préalable de la syntaxe formelle et de la sémantique des langages pour les ontologies.

Dans cet article, nous proposons IOPE (Interactive Ontology Population and Enrichment), un framework pour la construction automatique d'une interface utilisateur (IHM) composée de *pages Web pré-remplies*. Nous exploitons les pages Web comme un moyen d'interaction naturel pour relever le défi des représentations graphiques peu intuitives des ontologies. IOPE génère les pages Web à partir de *contraintes ontologiques*, supportant à la fois le processus de mise à jour contrôlée pour une ontologie donnée, et le préremplissage des pages générées à partir des instances de l'ontologie. Bien que IOPE soit générique et applicable à des ontologies de divers domaines, nous utilisons une ontologie appelée ONTOSAMSEI [2] comme cas d'utilisation dans lequel des experts sont guidés pour spécifier des ateliers de simulation de gestes médicaux. Dans [1] nous utilisons un ensemble d'expérimentations dans ce domaine pour montrer l'efficacité et l'efficience de notre approche. L'interface graphique IOPE de ONTOSAMSEI est accessible via le lien suivant : *http://iope.tabasi.info*. L'ontologie ONTOSAMSEI est disponible sur la plateforme Perscido de partage des données de la recherche : *https://perscido.univ-grenoble-alpes.fr/datasets/DS352*. Une version longue révisée de cet article publiée à WISE'21 est disponible sur HAL : *https://hal.archives-ouvertes.fr/hal-03671035*.

2 Présentation d'IOPE

Notre approche consiste à transformer la représentation des données RDF et des contraintes d'une ontologie de domaine sous forme d'une interface graphique interactive. Cette interface guide les experts du domaine dans l'exploration de l'ontologie et permet de modifier et enrichir cette ontologie via des composants interactifs. Toutes les interactions des experts dans IOPE sont traduites en triplets RDF qui sont contrôlés par un spécialiste en gestion de connaissance pour valider la cohérence de l'ontologie. Les différentes étapes d'IOPE sont brièvement présentés ci-dessous.

Input : Une ontologie de domaine est le point de départ pour construire l'interface graphique. Un algorithme de saturation, détaillé dans [1], s'emploie à saturer l'ensemble des contraintes de l'ontologie relative à une même hiérarchie de concepts.

IOPE GUI : L'interface graphique se compose d'un ensemble de pages Web liées et pré-remplies, générées auto-

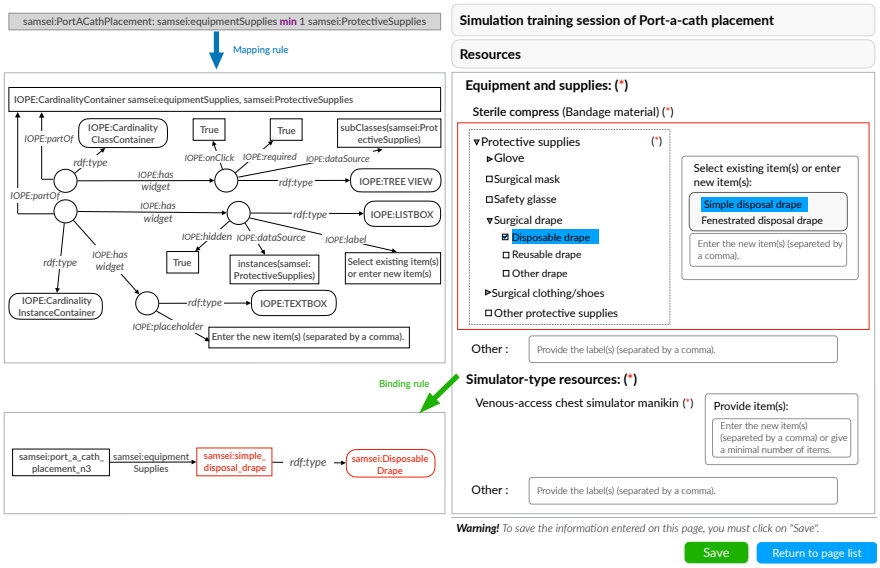


FIGURE 1 – Une exemple de page web pré-remplie avec ses règles de mapping et binding.

matiquement à partir des contraintes de l'ontologie. Un ensemble de 16 *règles de mapping* spécifie le changement de représentation. L'application de ces règles suit une logique de chainage avant. Un ensemble de 9 *règles de binding* permet de tisser des liens entre les formulaires Web et l'ontologie afin de générer un graphe RDF résultat des interactions de l'expert du domaine.

Output : Un ensemble de triplets RDF, résultat des interactions utilisateurs (ajouts et modifications des données).

La figure 1 montre l'une des pages Web préremplies (à droite) générées à partir des règles de mapping sur les contraintes (à gauche) et représente ici les ressources requises d'un atelier (comme l'équipement, les fournitures, et les types de simulateur). En haut à gauche de la figure, une contrainte ontologique au format Turtle , explicite qu'au moins un équipement de type "protection" est obligatoire pour cet atelier. Évidemment, le format Turtle est difficile à comprendre pour les utilisateurs sans connaissances OWL/RDFS. L'implémentation HTML est mise en évidence par un liseré rouge. Les interactions de l'utilisateur avec la page Web pré-remplie sont surlignées en bleu : le formateur sélectionne une sous-classe Disposable drape/Drap jetable dans la hiérarchie Protective supplies/Fournitures de protection, et choisit l'instance Drap jetable simple dans la liste d'instances fournie. Les interactions sont transformées en graphes RDF en utilisant la règle indiquée dans la partie inférieure gauche de la figure (binding rule). Cette transformation aboutit à deux triplets RDF : une relation entre une instance de l'atelier et l'instance sélectionnée, et une relation entre cette dernière et sa classe.

Références

[1] Shadi Baghernezhad-Tabasi et al. IOPE : Interactive ontology population and enrichment guided by ontological constraints. In *WISE*, 2021.

[2] Shadi Baghernezhad-Tabasi et al. OntoSAMSEI : Interactive ontology engineering for supporting simulation-based training in medicine. In *WETICE*, 2021.

Évaluation automatique d'alignements complexes : une approche basée sur des instances

Elodie Thiéblin, Ollivier Haemmerlé, Cassia Trojahn
Institut de Recherche en Informatique de Toulouse

elodie@thieblin.fr, ollivier.haemmerle@irit.fr, cassia.trojahn@irit.fr

Résumé

Un alignement d'ontologies est un ensemble de correspondances entre les entités de différentes ontologies. La plupart des travaux sur l'évaluation d'alignements se sont concentrés sur l'évaluation d'alignements simples (i.e., dont les correspondances sont entre une seule entité de l'ontologie source et une seule entité de l'ontologie cible). L'émergence d'outils d'alignement capables de générer des alignements complexes (i.e., dont des correspondances incluent des constructeurs logiques ou des fonctions de transformation) a fait apparaître le besoin d'outils d'évaluation automatique de ces alignements. Ce papier propose i) un système d'évaluation automatique d'alignements complexes fondé sur des requêtes et des instances, et ii) un jeu de données sur l'organisation de conférences. Ce jeu de données est composé d'ontologies peuplées et d'un ensemble de questions de compétence pour alignement sous la forme de requêtes SPARQL. Des alignements de l'état de l'art sont évalués sur le jeu de données et les difficultés sur l'évaluation d'alignements sont discutées.

Mots-clés

alignement d'ontologies, alignement complexe, évaluation, jeu de données d'évaluation

Abstract

Ontology matching is the task of generating a set of correspondences (i.e., an alignment) between the entities of different ontologies. While most efforts on alignment evaluation have been dedicated to the evaluation of simple alignments (i.e., those linking one single entity of a source ontology to one single entity of a target ontology), the emergence of matchers providing complex alignments (i.e., those composed of correspondences involving logical constructors or transformation functions) requires new strategies for addressing the problem of automatically evaluating complex alignments. This paper proposes i) a benchmark for complex alignment evaluation composed of an Automatique evaluation system that relies on queries and instances, and ii) a dataset about conference organisation. This dataset is composed of populated ontologies and a set of competency questions for alignment as SPARQL queries. State-of-the-art alignments are evaluated and a discussion on the difficulties of the evaluation task is provided.

Keywords

ontology matching, complex alignment, evaluation, benchmark

1 Introduction

Cet article est une version française de [1]. Un alignement d'ontologies est un ensemble de correspondances entre les entités de différentes ontologies. Cela sert de base pour de nombreuses tâches comme l'intégration de données, l'évolution d'ontologies ou la réécriture de requêtes. Bien que ce champ de recherche se soit bien développé, la plupart des travaux se concentrent sur la génération de correspondances simples (i.e., lier une entité d'une ontologie source à exactement une entité d'une ontologie cible). Toutefois, les correspondances simples ne permettent pas de couvrir toute l'hétérogénéité des ontologies à aligner. Les correspondances complexes expriment des relations plus expressives entre les ontologies. Par exemple, l'information selon laquelle un article a été accepté dans une conférence peut s'exprimer par une classe *ekaw:Accepted_Paper* ou par une restriction sur la propriété *cmt:hasDecision* sur la classe *cmt:Acceptance*. La correspondance $\langle$*ekaw:Accepted_Paper*, $\exists$*cmt:hasDecision.cmt:Acceptance*, $\equiv$, *1*$\rangle$ exprime l'équivalence entre les deux représentation d'un "article accepté" avec une confiance de 1.

Le besoin pour des alignements complexes a été décrit assez tôt [2, 3], et de nombreux outils de génération d'alignements complexes ont suivi, comme le présente l'état de l'art sur le sujet [4]. En revanche, peu d'initiatives se sont concentrées sur l'évaluation de ces alignements. La plupart des outils d'alignement complexe ont été évalués manuellement [5], en général seulement en termes de précision, ou sur des jeux de données spécifiques à leur outil [6], sur lesquels un rappel est calculé.

Bien que de nombreux efforts soient déployés sur l'évaluation automatique d'alignements, notamment dans les campagnes de l'Ontology Alignment Evaluation Initiative (OAEI) [1], la plupart se concentrent sur les alignements simples. Récemment, la première tâche d'alignement complexe a été proposée à l'OAEI [7], ouvrant de nouvelles perspectives à l'évaluation automatique d'alignements complexes.

1. http ://oaei.ontologymatching.org/

Dans cet articles, un jeu de données constitué d'ontologies aux instances contrôlées et partagées, ainsi qu'un ensemble de questions de compétences sous forme de requêtes SPARQL sont proposés, ainsi qu'un système d'évaluation automatique des alignements. Tandis que les outils et données d'évaluation d'alignements [8, 9] se fondent sur des alignements de référence et mesurent la similarité entre l'alignement évalué et le référent, nous proposons un ensemble de questions de compétences pour alignements (CQA) comme référence. Une CQA exprime, via une requête SPARQL, la connaissance qu'un alignement devrait couvrir entre les ontologies source et cible [10]. Nous proposons deux métriques d'évaluation, la *couverture de CQA*, fondée sur des paires de requêtes SPARQL équivalentes mesure à quel point l'alignement évalué couvre les besoins des requêtes ; la *précision intrinsèque* compare les instances des membres de la correspondance. La précision intrinsèque équilibre la couverture de CQA comme la précision équilibre le rappel.

Cet article

- discute des défis de l'évaluation automatique d'alignements complexes par rapport aux approches d'évaluation existantes ;
- propose une approche automatique pour évaluer l'alignement complexe ;
- propose un jeu de données d'ontologies peuplées de manière contrôlée et des questions de compétence pour alignement associées ;
- évalue des alignements de l'état de l'art sur le jeu de données et discute des résultats.

Le système et le jeu de données sont publiés sous licence LGPL [2].

2 Contexte

2.1 Alignement complexe d'ontologies

Un alignement A lie deux ontologies : une ontologie source o et une cible o' [11]. A est directionnel, noté $A_{o \to o'}$. $A_{o \to o'}$ est un ensemble de correspondances $\langle e, e', r, n \rangle$. Chaque correspondance exprime une relation r (e.g., equivalence ($\equiv$), subsomption ($\sqsupseteq$, $\sqsubseteq$)) entre ses deux membre e et e', et n exprime le degré de confiance $[0..1]$ dans cette correspondance. Un membre peut être une entité simple de l'ontologie (classe, propriété sur les données, sur les objets ou individu) de respectivement o et o' ou une construction plus complexe composée d'entités et de constructeurs ou de fonctions de transformation. Nous considérons deux types de correspondances fondées sur leurs membres [12] :

- une correspondance est **simple** si e et e' sont des entités simples (representés avec une IRI) : $\langle$*ekaw:Paper, cmt:Paper,* $\equiv$*, 1*$\rangle$
- une correspondance est **complexe** si e et/ou e' implique un constructeur ou une fonction de transformation $\langle$*ekaw:Accepted_Paper,* $\exists$ *cmt:hasDecision.cmt:Acceptance,* $\equiv$*, 1*$\rangle$ et $\langle$*concatenation(edas:hasFirstName, " ", edas:hasLastName), cmt:name,* $\to$*, 1*$\rangle$

Une correspondance simple est notée (s:s). Une correspondance complexe peut-être (s:c), si son membre source est une entité simple, (c:s) si son membre cible est une entité simple, ou (c:c) si ses deux membres sont complexes.

2.2 Questions de compétence pour alignement (CQA)

En conception d'ontologies, les questions de compétences (CQ) ont été introduites comme *les besoins en connaissance sous la forme de questions auxquels l'ontologie doit pouvoir répondre* [13]. Comme définie dans [10], une question de compétence pour alignement (CQA) est une question de compétence qui devrait être ouverte par deux ontologies ou plus, i.e., elle exprime la connaissance qu'un alignement devrait couvrir (si les deux ontologies permettent elles-mêmes d'y répondre).

La première différence entre une CQA et une CQ est que le champ d'une CQA est limité par l'intersection des champs des ontologies source et cible. La seconde est que ce champ maximal et idéal d'un alignement n'est pas connu a priori puisqu'il est le but même de l'alignement.

Comme pour les CQ [14], une CQA peut être exprimée en langage naturel ou comme une requête SPARQL SELECT. Inspirée de la notion d'arité de prédicat [14], l'*arité d'une CQA* représente l'arité des réponses attendues à une CQA [10] :

- Une question *unaire* attend un ensemble d'instances ou valeurs, e.g., "Quels sont les articles acceptés ?" *(paper1), (paper2)*.
- Une question *binaire* attend un ensemble de paires d'instances ou valeurs, e.g., "Quelle est la décision des articles ?" *(paper1, accept), (paper2, reject)*.
- Une question *n-aire* attend un tuple de taille n, e.g., "Quelle est la décision associée à la relecture des articles ?" *(paper1, review1, weak accept), (paper1, review2, reject)*.

3 État de l'art

L'évaluation de système d'alignement se fait sur un jeu de données d'évaluation, généralement composé d'un ensemble d'ontologies, d'un alignement de référence et potentiellement d'autres entrées variées (requêtes, instances, alignement partiel, etc.). L'alignement généré est évalué par un **outil d'évaluation** qui lui donne un score. Différentes dimensions d'évaluation peuvent être considérées :

Ressources Consommation de l'outil (mémoire, temps, CPU)

Entrées contrôlées Évaluation de l'outil en faisant varier ses paramètres d'entrée, comme dans les tâches GeoLink et Hydrography [7]

Sortie Évaluation de l'alignement lui-même, soit sur ses propriétés intrinsèques [15, 16], soit sur sa conformité à un alignement de référence.

Orienté tâche Évaluation de l'alignement sur son application à une tâche donnée [17, 18]

2. https://framagit.org/IRIT_UT2J/conference-dataset-population

3.1 Métriques d'évaluation

Les travaux ayant proposé des outils d'alignement complexe ont été évalués manuellement en termes de précision [5, 6, 19, 20], ou sur des jeux de données spécifiques (parfois manuellement créés) à leur outil pour calculer un rappel [6, 20].

Les métriques *accuracy* et *top*$-x$ *accuracy* [21] ont été appliquées dans des évaluations où le nombre de correspondances recherchées est prédéfini, e.g., une seule correspondance est attendue pour chaque entité de l'ontologie cible. L'accuracy est le pourcentage de "questions" (ou sous-tâches) prédéfinies ayant une réponse correcte. Une "question" dans ce contexte pourrait être une entité de l'ontologie cible à aligner et les "réponses" les correspondances ayant cette entité comme membre cible. Certains outils génèrent des réponses différentes pour chaque question, e.g., une liste ordonnée de correspondances pour chaque entité cible. Dans ce cas, la top$-x$ accuracy est le pourcentage de questions pour lesquelles la réponse correcte se trouve dans les x premières réponses à la question.

L'approche [22], pour évaluer les correspondances complexes entre ontologies agronomiques, se fonde sur une comparaison manuelle de requêtes de référence et de requêtes automatiquement réécrites grâce à l'alignement.

3.2 Approches et jeux de données

La première tâche d'alignement complexe de l'OAEI [7] consiste en quatre jeux de données sur des domaines variés avec des stratégies d'évaluation variées :

Complex conference Un alignement consensuel contenant des correspondances (s :c) a été créé, fondé sur la méthodologie de la réécriture de requête [12]. Chaque correspondance est manuellement classifiée comme vrai positif ou faux positif par rapport à l'alignement de référence.

Hydrography and GeoLink Ce jeu porte sur des ontologies sur l'hydrographie et sur la GeoScience [23]. Les alignements sont évalués sur les tâches suivantes : i) trouver toutes les entités qui apparaissent dans une correspondance, ii) trouver la construction correcte reliant les entités et iii) trouver les correspondances complexes à partir de rien. En 2018, uniquement la première tâche a été implémentée [24]. En 2019, une métrique proche de la précision et rappel relaxés [25] a été appliquée aux tâches i et ii.

Taxon un ensemble de CQA sur des bases de connaissances agronomiques sont réécrites en utilisant l'alignement évalué. Chaque requête réécrite est classifiée manuellement comme sémantiquement équivalente ou non. Chaque correspondance est également manuellement classifiée comme vrai positif ou faux positif sans référence.

3.3 Évaluation orientée tâche

Certaines applications des alignements comme l'évolution d'ontologie ou la réponse aux questions (query answering) ont des contraintes différentes en termes de couverture et de temps d'exécution [11]. La tâche *OA4QA* [26] s'est spécialisée sur la réponse aux questions. Cette tâche a utilisé une version artificiellement peuplée du jeu de données *Conférence* et un ensemble de requêtes manuellement créées sur ces instances. Une requête exprimée avec l'ontologie *cmt* est exécutée sur l'ontologie fusionnée *cmt* $\cup$ *ekaw* $\cup$ *A*, où *A* est l'alignement entre *cmt* et *ekaw*. Les résultats de la requête étaient comparés à ceux sur l'ontologie *cmt* $\cup$ *ekaw* $\cup$ *ra1*, *ra1* étant l'alignement de référence.

[27] propose une évaluation "end-to-end" de requêtes réécrites à partir d'un alignement évalué. Les résultats des requêtes sont manuellement classifiés sur une échelle à 6 points par rapport à leur pertinence pour un utilisateur. Cette évaluation a été menée avec deux systèmes de réécriture de requêtes. Si un membre source e n'apparaît pas dans les correspondances, le système "vers le haut" cherchera une classe parente de e dans les membres sources et le système "vers le bas" cherchera une classe enfant de e.

La réécriture de requêtes est une des applications importantes pour l'alignement complexe. Évaluer de tels alignements sur cette tâche est par conséquent pertinent. La réécriture de requêtes fondée sur des alignements simples peut se contenter d'une approche naïve consistant à remplacer l'IRI d'un membre source par celle du membre cible dans la requête [28]. La tâche n'est pas aussi aisée pour les alignements complexes où la sémantique de l'alignement elle-même doit être prise en considération. [29] présente une approche de réécriture pour des requêtes SPARQL CONSTRUCT spécifiques mais la plupart des systèmes de réécriture de requêtes se fondent sur des correspondances simples ou (s :c) et échouent à gérer les correspondances (c :c) très expressives.

3.4 Positionnement

Pour l'évaluation des alignements complexes, des travaux se concentrent sur une évaluation manuelle en termes de précision [5, 19], calculer le rappel sur des patrons récurrents entre ontologies [6, 20], ou se fondent sur un échantillon de correspondances de référence [30]. Tandis que ces approches se concentrent sur la comparaison de correspondances, nous transférons le problème sur de la comparaison d'instances. Nous proposons une évaluation qui considère des requêtes comme référence et dont les métriques sont la couverture de requêtes (comme rappel) et la précision intrinsèque (comme la précision, mais sans alignement de référence). Par conséquent, notre approche nécessite des jeux de données peuplées de manière contrôlée.

Comme [26], la référence pour l'évaluation est un ensemble de requêtes et non un alignement. Proche de notre approche, [26] se fonde sur le jeu de données *Conférence* peuplé artificiellement. Toutefois, leurs requêtes sont exécutées sur une ontologie fusionnée et leurs alignements sont limités aux correspondances simples. Dans notre cas, les requêtes sont exécutées sur les différentes ontologies peuplées. Similairement à [27], les requêtes sont réécrites automatiquement avec l'alignement évalué. Toutefois, notre approche est entièrement automatique et ne se fonde pas sur une classifica-

tion manuelle des requêtes réécrites.
La Table 1 résume les approches d'évaluation d'alignements complexes proches de notre proposition (**CQA** benchmark, en gras dans la Table 1).

4 Évaluation automatique d'alignements complexes

La plupart des métriques d'évaluation pour les alignements simples ou complexes se fondent sur une comparaison syntaxique ou sémantique, mais très peu sur la comparaison au niveau des instances.
La comparaison *syntaxique* des alignements mesure l'effort à fournir pour transformer une correspondance évaluée en celle de référence. Toutefois, cela ne prend pas en compte le cas où une correspondance est sémantiquement équivalente à celle de référence mais utilise des constructeurs ou des niveaux de factorisation différents. Une comparaison syntaxique dépend également du langage et de la manière dont les correspondances sont exprimées. Par exemple, $\langle$ *o:Author* , $\exists o'$*:authorOf.*$\top$, $\equiv \rangle$ est sémantiquement équivalent à $\langle$ *o:Author* , $\exists o'$*:writtenBy*$^{-}$.$\top$, $\equiv \rangle$ mais ces correspondances utilisent des IRI et des constructeurs différents et sont donc syntaxiquement différentes. Un problème de factorisation reviendrait à comparer $\langle$ *o:paperWrittenBy , dom(*o'*:Paper)* $\sqcap$ o'*:writes*$^{-}$, $\equiv \rangle$ et $\langle$ *o:paperWrittenBy , (*o'*:writes* $\sqcap$ *range(*o'*:Paper))*$^{-}$, $\equiv \rangle$ qui sont deux correspondances équivalentes. Le constructeur *inverse* est factorisé dans la seconde correspondance. La comparaison syntaxique de requêtes fait face au même problème : des requêtes syntaxiquement différentes peuvent être sémantiquement équivalentes.
La comparaison *sémantique* compare le sens des formules. La précision et le rappel sémantiques comparent l'ensemble des axiomes inférés de la fusion des ontologies avec l'alignement évalué à la fusion des ontologies avec l'alignement de référence. Pour les alignements complexes, cela a l'avantage que tout élément traduisible en OWL puisse être évalué de cette manière. Toutefois, l'expressivité de l'alignement évalué fusionné avec les ontologies est limité à $\mathcal{SROIQ}$ (le fragment décidable de OWL [31]). Les correspondances avec des fonctions de transformation ne peuvent pas être comparées de cette manière non plus. La comparaison sémantique de requêtes proposée par [32] se fonde sur l'imbrication de requêtes, qui peut nécessiter des inférences. Cette comparaison ne peut également pas s'appliquer aux fonctions de transformation.
La comparaison *fondée sur les instances* (de correspondances ou de résultats de requêtes) est une alternative aux deux méthodes sus-mentionnées. Elle a l'inconvénient de nécessiter que les ontologies soient peuplées et de manière contrôlée.
Nous proposons deux métriques d'évaluation. La *couverture de CQA* (ou CQA Coverage) mesure à quel point un alignement permet de traduire un ensemble de CQA. La *précision intrinsèque* (ou intrinsic precision) compare les instances des membres des correspondances. La précision intrinsèque équilibre la couverture de CQA comme la précision équilibre le rappel en recherche d'information.

4.1 Déroulé d'évaluation

Figure 1 présente le déroulé d'évaluation adopté par notre approche. Les étapes du déroulé sont :

1. **Sélection Ancre** Cette étape consiste à renvoyer une paire d'objets comparables $\langle x_i, x_{rj} \rangle$. x_i est un objet lié à A_{eval} and x_{rj} lié à $reference$. Dans le cas où la référence est une requête ou une paire de requêtes équivalentes, x_i peut être une requête dérivée de A_{eval} et x_{rj} une requête de référence.
2. **Comparaison** L'étape de comparaison a pour but de renvoyer une relation $rel(x_i, x_{rj})$ pour chaque paire obtenue précédemment $\langle x_i, x_{rj} \rangle$. La relation peut être une équivalence (*i.e.*, $x_i \equiv x_{rj}$), une subsomption, une intersection, une disjonction, *etc.* Une valeur de similarité peut être associée à la relation. Dans notre approche, la comparaison est fondée sur les instances.
3. **Score** Cette étape associe un score à chaque relation trouvée précédemment $rel(x_i, x_{rj})$.
4. **Agrégation** Les scores sont localement et globalement agrégés pour obtenir le $score\ final$. Les agrégations peuvent être faites par : meilleur candidat, moyenne, moyenne pondérée, *etc.* L'agrégation locale agrège les scores pour un objet donné. Il peut y avoir différentes agrégations locales. Par exemple, il peut y avoir une agrégation par rapport à l'objet évalué et par rapport à l'objet de référence. L'agrégation globale se fait sur les scores localement agrégés pour obtenir le score final.

4.2 Couverture de CQA

La référence est un ensemble de CQA équivalentes sous la forme de requêtes SPARQL. Un alignement évalué A_{eval} sera utilisé pour réécrire chaque CQA source. La requête réécrite sera comparée avec la requête CQA cible. La comparaison des requêtes est fondée sur les instances et une valeur est associée à la relation entre les deux requêtes fondée sur l'intersection des instances renvoyées la requête évaluée et par la CQA cible. Différentes fonctions de scores sont appliquées en fonction de la relation entre requêtes. Une agrégation de meilleur candidat est faite localement sur les requêtes de référence. Une moyenne des scores localement agrégés est faite pour obtenir le score final.

4.2.1 Ancrer la CQA source

La référence dans cette évaluation est un ensemble de CQA sous la forme de requêtes SPARQL sur les ontologies cible et source. Chaque CQA source cqa_s a un équivalent cible cqa_t. Dans l'étape de sélection d'ancre, chaque cqa_s est réécrit en utilisant l'alignement évalué A_{eval}. L'étape de réécriture renvoie toutes les requêtes cibles possibles générées par le système de réécriture à partir de A_{eval} dans l'ensemble Q_t. Pour chaque requête q_t dans Q_t, une paire (q_t, cqa_t) est formée.
Deux systèmes de réécritures ont été considérés. Aucun de ces système ne considère la relation d'une correspondance

TABLE 1 – Comparaison des approches d'évaluation d'alignements complexes. Le *Type de corresp.* la forme la plus expressive que permet de gérer l'approche – (c :c) est plus complexe que (s :c), qui est plus complexe que (s :s).

Benchmark	*Type d'évaluation*	*Type de référence*	*Type de corresp.*
OA4QA [26]	Automatique (precision/recall)	Requête	(s :s)
Query rewrite [27]	Manuelle	Requête	(s :s)
Patterns evaluation [6]	Manuelle	Alignement	(s :c)
Patterns evaluation [20]	Manuelle	Alignement	(s :c)
Thieblin 2018 [12]	Manuelle	Alignement	(s :c)
GeoLink 2018 [23]	Automatique (precision/recall) /Manuelle	Alignement	(c :c)
Hydrography 2018 [23]	Automatique (precision/recall)/Manuelle	Alignement	(c :c)
GeoLink 2019	Automatique (relaxed precision/recall)	Alignement	(c :c)
Hydrography 2019	Automatique (relaxed precision/recall)	Alignement	(c :c)
Taxon [22]	Manuelle	Requête	(c :c)
CQA benchmark	**Automatique basée sur instances (couverture de CQA/ précision intrinsèque)**	**Requête**	**(c :c)**

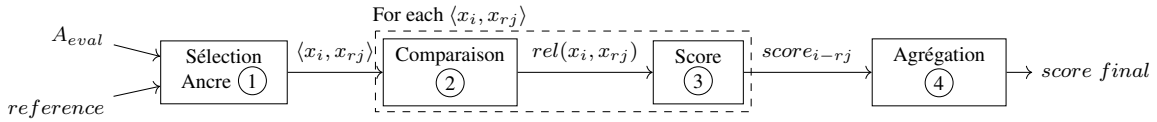


FIGURE 1 – Déroulé de l'évaluation de l'alignement A_{eval} avec une $reference$ générique.

ni la valeur de confiance lui étant attribuée. Le premier système [33] réécrit chaque triplet du patron de graphe de cqa_s en utilisant A_{eval}. Quand le prédicat ou l'objet du triplet apparaît comme membre source d'une correspondance dans A_{eval}, le membre cible de la correspondance est transformé en patron de graphe SPARQL et remplace le triplet dans la requête. Ce système ne gère que les correspondances (s :c). Si un triplet peut être réécrit avec plusieurs correspondances, toutes les combinaisons possibles sont ajoutées à Q_t. Par exemple,

```
Q1 = SELECT ?s WHERE{?s a
ekaw:Accepted_Paper.}
```

contient *ekaw:Accepted_Paper*, membre source de $c_1 =$ ⟨*ekaw:Accepted_Paper,* ∃*cmt:hasDecision.cmt:Acceptance,* ≡*, 1*⟩.

La requête réécrite en utilisant la correspondance c_1 :

```
Q2 = SELECT ?s WHERE{?s cmt:hasDecision
?o. ?o a cmt:Acceptance.}
```

Ce système de réécriture ne peut malheureusement par fonctionner dans l'autre sens : `Q2` ne pourra pas être réécrite avec c_1.

Le second système est fondé sur les instances et a été développé dans le contexte de cet article. Les instances I_s^{cqa} de cqa_s sont récupérées de l'ontologie source. Pour chaque correspondance c de A_{eval}, le membre source est transformé en requête pour récupérer les instances I_s de l'ontologie source. Si $I_s \equiv I_s^{cqa}$, alors le membre cible de c est transformé en requête et ajouté à Q_t.

Par exemple, `Q1` récupère un ensemble d'instances de papier accepté de l'ontologie *ekaw*. Cet ensemble d'instances est comparé aux membres sources de chaque correspondance de A_{eval}. Dans ce cas, *ekaw:Accepted_Paper* décrit les mêmes instances que le membre source de c_1. Son membre cible peut être traduit en requête SPARQL et devient `Q2`.

Ce système permet de réécrire des requêtes dans les deux sens, donc de gérer les correspondances (c :c). `Q2` pourra être réécrite avec l'inverse de c_1 (l'inverse d'une correspondance est son équivalent où le membre source devient le membre cible et vice-versa). Toutefois, ce système ne permet pas de combiner les correspondances entre elles.

Les deux systèmes de réécriture de requêtes sont utilisés.

4.2.2 Comparaison

Les instances I_t^{cqa} de cqa_t et I_t de q_t sont récupérées de l'ontologie cible. I_t et I_t^{cqa} sont comparés. Nous calculons deux scores par cette comparaison : précision de requête (QP) et de rappel de requête (QR). Ces valeurs sont associées à la relation $rel(q_t, cqa_t)$.

$$QP = \frac{|I_t \cap I_t^{cqa}|}{|I_t|} \qquad QR = \frac{|I_t \cap I_t^{cqa}|}{|I_t^{cqa}|}$$

$$rel(q_t, cqa_t) = \begin{cases} \equiv \text{ if } QR = 1 \text{ and } QP = 1 \\ \sqsubseteq \text{ if } QR \leq 1 \text{ and } QP = 1 \\ \sqsupseteq \text{ if } QR = 1 \text{ and } QP \leq 1 \\ overlap \text{ if } 0 < QR \leq 1 \text{ and} \\ 0 < QP \leq 1 \\ \perp \text{ if } QR = 0 \text{ and } QP = 0 \end{cases}$$

4.2.3 Score

Les valeurs de précision de requête et rappel de requête entre cqa_t et q_t sont transformées en un score de F-mesure de requête par une moyenne harmonique.

$$Fmesure = 2 \times \frac{QR \times QP}{QR + QP}$$

La F-mesure de requête équilibrant précision et rappel de requête a été choisie par rapport aux autres métriques (classique, relaxée).

4.2.4 Agrégation

Comme l'étape de réécriture renvoie toutes les requêtes possibles sans prendre en compte la relation de la correspondance, du bruit est introduit. De plus, la même requête peut être générée par les deux systèmes de réécriture. Par conséquent, pour chaque cqa_t, la requête q_t ayant le meilleur score est gardée. Cette agrégation par meilleur candidat permet de limiter l'impact du bruit introduit par l'étape de réécriture. Si cqa_s n'a pas pu être réécrit par l'alignement, son score est 0. L'agrégation globale est faite par moyenne sur les scores de chaque CQA.

4.3 Précision intrinsèque

La couverture de CQA agrège les résultats par CQA et non par requête réécrite à cause du bruit introduit par les systèmes de réécriture de requêtes. Cela induit qu'un alignement ayant toutes les correspondances possibles (correctes et erronées) entre une ontologie source et cible obtiendrait un bon score de couverture de CQA. Pour contrebalancer cela, nous proposons de mesurer la précision intrinsèque fondée sur les instances d'un alignement. Pour chaque correspondance c_i de A_{eval}, les instances I_s représentées par le membre source de la correspondance sont comparées aux instances I_t représentées par le membre cible. Chaque correspondance est classifiée comme *équivalence*, *sumbsomption*, *intersection*, *disjonction* en fonction de la relation entre I_s et I_t, ou *vide* if $I_s = I_t = \emptyset$. Une correspondance peut être *vide* si ses deux membres sont des entités insatisfiables ou des entités non peuplées.
Des scores de précision sont donnés pour chaque classe de correspondance : *équivalence* mesure le pourcentage de correspondances dont les membres sont exactement peuplés avec les mêmes instances, *subsomption* celles dont un membre subsume l'autre, *intersection* celles ayant une intersection non nulle et *non-disjointes* considère toutes les correspondances sauf les *disjonctions*.

5 Jeu de données fondé sur des CQA

5.1 Méthodologie de création du jeu de données

Le but est de créer un jeu de données sur lequel les outils d'alignement d'ontologies peuvent fonctionner et sur lequel l'évaluation décrite dans la section précédente peut être exécutée. Ce jeu de données doit donc contenir des ontologies peuplées ainsi qu'un ensemble de CQA exprimées sous forme de requêtes SPARQL sur ces ontologies.
Nous proposons la méthodologie suivante :

1. Créer un ensemble de CQA fondé sur un scénario applicatif. Seules les CQA unaires et binaires sont considérées dans ce travail.
2. Créer un format pivot (qui servira de pierre de rosette entre les différentes ontologies) couvrant toutes les CQA précédemment définies.
3. Pour chaque ontologie du jeu de données, créer des requêtes SPARQL INSERT correspondant au format pivot.
4. Instancier le format pivot avec des données réelles ou artificielles.
5. Peupler les ontologies avec le format pivot instancié en utilisant les requêtes SPARQL INSERT.
6. Faire tourner un raisonneur pour vérifier la consistance des ontologies peuplées. Si une erreur est levée, changer son interprétation de l'ontologie et itérer sur les point 3 à 5.
7. En se fondant sur les requêtes SPARQL INSERT, traduire les CQA couvertes par deux ontologies ou plus comme requêtes SPARQL SELECT.

Dans cette méthodologie, l'interprétation des ontologies est identique pour le peuplement et pour la création des CQA. La création des CQA peut être faite avec des utilisateurs et des experts du domaine, comme recommandé dans la méthodologie NeOn [34] pour les questions de compétence. Les CQA peuvent aussi dériver des questions de compétence utilisées pour concevoir les ontologies du jeu de données. Dans notre implémentation, un expert a créé les CQA, et elles ont été validées par un second expert qui a jugé qu'elles étaient assez exhaustives pour couvrir un scénario d'organisation de conférence.

5.2 Jeu de données conférence

Le jeu de données conférence [35] a été largement utilisé [9], en particulier lors des campagnes de l'OAEI. Il se compose de 16 ontologies sur le domaine de l'organisation de conférences et d'un alignement de référence simple entre 7 d'entre elles. Ces ontologies ont été développées individuellement. Des alignements complexes de référence ont été proposés pour 5 ontologies de ce jeu de données [12]. Dans la première tâche complexe de l'OAEI, une évaluation a été proposée sur un alignement complexe consensuel entre 3 ontologies (*cmt*, *conference*, *ekaw*) [7].
Nous avons peuplé les 5 ontologies des alignements de référence de [12] : *cmt*, *conference* (Sofsem), *confOf* (confTool), *edas* et *ekaw* (Table 2).

TABLE 2 – Number of entities by type of each ontology.

	cmt	conference	confOf	edas	ekaw
Classes	30	60	39	104	74
Obj. prop.	49	46	13	30	33
Data prop.	10	18	23	20	0

Bien que ce jeu de données ait été beaucoup utilisé, il a seulement été partiellement peuplé. Dans la tâche OA4QA, les classes sur lesquelles portaient les 18 requêtes de l'évaluation ont été peuplées mais la création de ces instances n'a pas été documentée.

5.3 Peuplement des ontologies de conférence

Un scénario d'organisation de conférence a été envisagé en examinant un cas réel : l'édition 2018 de la conférence ESWC. La liste de CQA en résultant a été étendue en explorant le champ couvert par les ontologies de conférence. Le format pivot a été instancié une première fois avec les données du site Web d'ESWC. L'étape du raisonneur a levé plusieurs erreurs comme la relation *cmt:hasAuthor* qui est fonctionnelle et ne représente donc que le premier auteur d'un article. Ayant identifié ces erreurs dans notre interprétation, nous avons repris la méthodologie étape par étape.

Deux erreurs qui n'ont pu être résolues par un changement d'interprétation ont nécessité une légère modification des ontologies.

— *cmt* : la relation *cmt:acceptPaper* entre un *Administrator* et un *Paper* était définie fonctionnelle et inverse fonctionnelle. Cela levait une erreur quand un administrateur acceptait plus d'un article. Elle a été modifiée pour n'être plus que inverse fonctionnelle.
— *conference* : *conference:Contribution_1st_author* était disjointe de *conference:Contribution_co-author*, ce qui levait une incohérence lorsqu'une personne était à la fois autrice d'un article et co-autrice d'un autre. La disjonction a été retirée.

Une ontologie n'est couverte que par CQA. Cela peut résulter en des populations hétérogènes pour certains concepts pourtant équivalents. Par exemple, *ekaw* et *cmt* ont toutes deux une classe *Document* mais la question *"Quels sont les documents ?"* n'est pas une CQA tandis que leurs sous-classes *Paper, Review, WebSite, Proceedings* sont le focus de CQA. Tandis que *ekaw* comporte les 4 sous-classes, *cmt* n'a que *Paper* et *Review*. La classe *cmt:Document* ne contiendra donc pas de site Web ou de proceedings tandis que *ekaw:Document* si.

Pour être proches de la réalité dans notre jeu de données, nous avons analysé les propriétés d'ISWC 2018 et ESWC 2017 à partir de Scholarly Data en complément d'ESWC 2018 et extrait des statistiques.

La première instanciation du format pivot à partir du site Web manquait d'informations pour représenter l'organisation d'une conférence. La liste des CQA a été étendue pour couvrir ce scénaria. En ont résulté une extension du format pivot et des requêtes SPARQL INSERT. Des données générées artificiellement à partir des statistiques des 3 conférences ont instancié le format pivot pour des conférences plus ou moins grandes (plus une conférence est grande, plus elle a de papiers soumis, de personnes, de membres du program committee, de workshop, etc.).

Pour permettre aux outils d'alignement fondés sur des statistiques d'être testés sur ce jeu de données, nous avons peuplé les mêmes ontologies avec des instances se recoupant plus ou moins. Chaque ontologie est peuplée avec plusieurs conférences, chaque conférence n'ayant aucune instance commune avec les autres. Cela permet de quantifier la partie d'instances communes à deux ontologies peuplées. 6 jeux de données en résultent, peuplés avec 25 conférences artificielles :

— 0 % : 5 conférences différentes par ontologie
— 20 % : 1 conférence commune pour toutes les ontologies et 4 différentes par ontologie
— ...
— 100 % : 5 conférences communes pour toutes les ontologies

Le pourcentage donné dans le nom des jeux de données est le pourcentage de conférences communes par ontologie. Comme la taille de chaque conférence est différente, le pourcentage des autres instances (articles, personnes, etc.) n'y sera pas identique.

5.4 CQA pour évaluation

Pour l'évaluation d'alignements, seules les CQA pouvant être couvertes par deux ontologies ou plus sont intéressantes. Nous avons élagué la liste des CQA pour enlever :

— les CQA couvertes par une seule ontologie
— les CQA se traduisant de la même manière pour toutes les ontologies (e.g., label d'une instance avec *rdfs :label*).

Table 3 présente le nombre de CQA initiales couvertes par chaque ontologie et ayant été gardées pour l'évaluation. 278 requêtes SPARQL SELECT en résultent.

TABLE 3 – Nombre de CQA initiales et d'évaluation couvertes par chaque ontologie

	cmt	conference	confOf	edas	ekaw	total
init.	46	90	67	60	84	152
eval.	34	73	54	52	65	100

6 Évaluation

Nous avons évalué 5 alignements avec notre approche. Nous indiquons le nombre de paires d'ontologies sur les 10 de notre jeu de données qui sont couvertes par ces alignements. Ces alignements ne contiennent pas de correspondances (c :c).

Query_rewriting [12] créé manuellement, contient 431 correspondances dont 191 complexes sur les 10 paires.

Ontology_merging [12] créé manuellement, contient 313 correspondances dont 54 complexes sur les 10 paires.

ra1 l'alignement simple de référence du jeu de données *conférence* [9] sur les 10 paires.

Ritze_2010 généré automatiquement [19], sur 4 paires. Il ne contient qu'une seule correspondance par paire.

Faria_2018 généré automatiquement [36] entre 3 paires.

L'alignement ra1 a été utilisé comme entrée pour Ritze_2010 et Faria_2018. Ra1 donc été ajouté à ces alignements pour le calcul de couverture de CQA. La couverture de CQA a été calculée sur tous les jeux de données pour calculer la déviation standard des scores de précision

de requête, rappel de requête et F-mesure de requête. La déviation standard est inférieure à 2×10^{-3} sur les 6 jeux de données et les 5 alignements.

Table 4 présente les résultats de l'évaluation sur le jeu de données 100%. Ritze_2010 et Faria_2018 ont une meilleure couverture que ra1 qu'ils incluent. Les correspondances qu'ils contiennent sont un complément aux correspondances simples pour couvrir les CQA. ra1 a une meilleure précision d'équivalence (0.56) que les autres alignements créés manuellement car il contient uniquement des correspondances avec des équivalences. Ce score est bas pour un alignement de référence pour la raison évoquée §5.3. Au vu des problèmes de peuplement, les scores d'intersection et non-disjointes donnent un bon aperçu de la précision d'un alignement. Nous avons choisi d'agréger la couverture de CQA et les scores de précisions d'intersection et non-disjointes respectivement en une moyenne harmonique (MH).

TABLE 4 – Couverture de CQA(C), précision intrinsèque(P), moyenne harmonique (MH)

	Query rew.	Onto. merg.	ra1	Faria 2018	Ritze 2010
C. CQA	**0.69**	0.63	0.42	0.41	0.48
P. équi.	0.42	0.43	0.56	0.65	0.75
P. subs.	0.80	0.80	0.83	0.71	0.75
P. inter.	0.90	0.86	**0.92**	0.71	0.75
P. non-dis.	0.94	0.91	**0.96**	0.71	0.75
MH inter.	**0.78**	0.73	0.58	0.52	0.59
MH non-dis.	**0.80**	0.74	0.58	0.52	0.59

Les alignements Query_rewriting et Ontology_merging obtiennent les meilleurs résultats, ce qui conforte leur rôle d'alignement complexe de référence sur ce jeu de données. Bien que ra1 obtienne la meilleure précision, sa couverture de CQA faible (0.42) montre que nombre de CQA nécessitent des alignements pour être traduites dans ce jeu de données. Faria_2018 et Ritze_2010 ne s'appliquent pas au même nombre de paires d'ontologies et ne peuvent donc pas être comparés aux autres.

Dans les résultats de l'OAEI 2018 [24], la précision mesurée pour Faria_2018 était de 0.54 (cf. Table 5). La précision intrinsèque donne les mêmes résultats pour la paire *cmt-ekaw*. Pour les autres paires, la différence est significative.

Pour la paire *conference-ekaw*, $\langle \exists$*conference:was_a_track-workshop_chair_of. conference:Tutorial, ekaw :Tutorial_Chair,* $\equiv$*, 0.369*$\rangle$ était considéré correct dans l'évaluation OAEI 2018. Pourtant, un axiome de l'ontologie *conference* restreint le domaine de *conference:was_a_track-workshop_chair_of* à *conference:Track* $\sqcup$ *conference:Workshop*. Cela a été pris en compte dans le peuplement de l'ontologie et la correspondance a été évaluée comme étant disjointe par le système d'évaluation.

TABLE 5 – Comparaison des scores de précisoin de Faria_2018 dans l'évaluation OAEI 2018 et basée sur nos métriques.

pair	OAEI 2018	P. equiv.	P. non-disj.
cmt-conference	0.4	1.00	1.00
cmt-ekaw	0.86	0.86	0.86
conference-ekaw	0.36	0.09	0.27
Average	0.54	0.65	0.71

7 Conclusions et perspectives

Nous avons présenté une approche d'évaluation automatique d'alignements complexes et son jeu de données associé. Par rapport à l'évaluation d'alignements simple, il est difficile de comparer les membres d'une correspondance évaluée à celle d'un alignement de référence. Tandis que les métriques d'évaluation fondées sur de la comparaison syntaxique échoueraient à couvrir l'ensemble des combinaisons possibles de constructeurs et IRI, les approches sémantiques restreignent l'expressivité des correspondances et des alignements à celle supportée par les raisonneurs, délaissant notamment les fonctions de transformation. La comparaison fondée sur les instances permet un compromis. Notre proposition déplace le problème sur la comparaison d'instances dans une tâche de réécriture de requête ciblant des besoins utilisateur. Nous avons proposé deux métriques d'évaluation. La couverture de CQA mesure à quel point un alignement permet de traduire un ensemble de requêtes SPARQL et la précision intrinsèque compare les instances des membres d'une correspondance. La précision intrinsèque équilibre la couverture de CQA de la manière dont la précision équilibre le rappel en recherche d'information.

Un jeu de données artificiel sur les ontologies de conférence a été proposé. Son peuplement a été contrôlé et guidé par des CQA.

Les systèmes de réécriture de requêtes sont limités aux correspondances (s :c), gérer les correspondances (c :c) reste une question ouverte. Nous avons proposé un système permettant de les prendre en compte seulement individuellement et non de les combiner. La réécriture de requêtes fondée sur les instances pourrait toutefois être une nouvelle piste à explorer dans cette voie. Nous n'avons pas évalué l'impact des systèmes de réécriture de requêtes sur l'évaluation, mais plutôt tenté de limiter leur impact avec une sélection du meilleur candidat.

L'approche proposée a été appliquée à des alignements existants et dans la campagne OAEI 2019. Notre approche d'évaluation attribue une bonne précision aux alignements de référence et une meilleure couverture de CQA pour les alignements complexes que simples. Toutefois, l'évaluation manuelle dans l'OAEI et celle proposée ici présentent des différences significatives.

L'évaluation d'alignements complexes et un défi trop large pour être résolu avec une seule approche et d'autres pistes, plus sémantiques, sont notamment à explorer.

Références

[1] É. Thiéblin, O. Haemmerlé, and C. Trojahn, "Automatic evaluation of complex alignments : An instance-based approach," *Semantic Web*, vol. 12, no. 5, pp. 767–787, 2021.

[2] P. R. S. Visser, D. M. Jones, B. T. J. M. Capon, and M. J. R. Shave, "An analysis of ontological mismatches : Heterogeneity versus interoperability," in *AAAI 1997 Spring Symposium on Ontological Engineering*, Stanford, USA, 1997, pp. 164–72.

[3] A. Maedche, B. Motik, N. Silva, and R. Volz, "Mafra — a mapping framework for distributed ontologies," in *Knowledge Engineering and Knowledge Management : Ontologies and the Semantic Web*, A. Gómez-Pérez and V. R. Benjamins, Eds. Berlin, Heidelberg : Springer Berlin Heidelberg, 2002, pp. 235–250.

[4] É. Thiéblin, O. Haemmerlé, N. Hernandez, and C. Trojahn, "Survey on complex ontology matching," *Semantic Web*, vol. 11, no. 4, pp. 689–727, 2020.

[5] D. Ritze, C. Meilicke, O. Sváb-Zamazal, and H. Stuckenschmidt, "A pattern-based ontology matching approach for detecting complex correspondences," in *4th International Workshop on Ontology Matching (OM-2009) collocated with the 8th International Semantic Web Conference (ISWC-2009)*, ser. CEUR Workshop, P. Shvaiko, J. Euzenat, F. Giunchiglia, H. Stuckenschmidt, N. F. Noy, and A. Rosenthal, Eds., vol. 551, 2009.

[6] B. Walshe, R. Brennan, and D. O'Sullivan, "Bayesrecce : A bayesian model for detecting restriction class correspondences in linked open data knowledge bases," vol. 12, no. 2, p. 25–52, Apr. 2016.

[7] É. Thiéblin, M. Cheatham, C. T. dos Santos, O. Zamazal, and L. Zhou, "The first version of the OAEI complex alignment benchmark," in *ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks co-located with 17th International Semantic Web Conference ISWC 2018*, ser. CEUR Workshop, M. van Erp, M. Atre, V. López, K. Srinivas, and C. Fortuna, Eds., vol. 2180, 2018.

[8] J. Euzenat, M. Rosoiu, and C. Trojahn, "Ontology matching benchmarks : Generation, stability, and discriminability," *Journal of Web Semantics*, vol. 21, pp. 30–48, 2013.

[9] O. Zamazal and V. Svátek, "The Ten-Year OntoFarm and its Fertilization within the Onto-Sphere," *Journal of Web Semantics*, vol. 43, pp. 46–53, 2017.

[10] É. Thiéblin, "Do competency questions for alignment help fostering complex correspondences ?" in *EKAW Doctoral Consortium 2018*, ser. CEUR Workshop, L. Hollink and F. Osborne, Eds., vol. 2306, 2018.

[11] J. Euzenat and P. Shvaiko, *Ontology Matching*, 2nd ed. Springer Berlin Heidelberg, 2013.

[12] É. Thiéblin, O. Haemmerlé, N. Hernandez, and C. Trojahn, "Task-oriented complex ontology alignment : Two alignment evaluation sets," in *The Semantic Web - 15th International Conference, ESWC 2018*, ser. Lecture Notes in Computer Science, A. Gangemi, R. Navigli, M. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, and M. Alam, Eds., vol. 10843. Springer, 2018, pp. 655–670.

[13] M. Grüninger and M. S. Fox, "Methodology for the design and evaluation of ontologies," in *Workshop on Basic Ontological Issues in Knowledge Sharing*, vol. 15, 1995.

[14] Y. Ren, A. Parvizi, C. Mellish, J. Z. Pan, K. van Deemter, and R. Stevens, "Towards competency question-driven ontology authoring," in *The Semantic Web : Trends and Challenges - 11th International Conference, ESWC 2014*, ser. Lecture Notes in Computer Science, V. Presutti, C. d'Amato, F. Gandon, M. d'Aquin, S. Staab, and A. Tordai, Eds., vol. 8465. Springer, 2014, pp. 752–767.

[15] C. Meilicke and H. Stuckenschmidt, "Incoherence as a basis for measuring the quality of ontology mappings," in *3rd International Workshop on Ontology Matching*, ser. CEUR Workshop, P. Shvaiko, J. Euzenat, F. Giunchiglia, and H. Stuckenschmidt, Eds., vol. 431, 2008.

[16] A. Solimando, E. Jiménez-Ruiz, and G. Guerrini, "Detecting and correcting conservativity principle violations in ontology-to-ontology mappings," in *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference*, ser. Lecture Notes in Computer Science, P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. A. Knoblock, D. Vrandecic, P. Groth, N. F. Noy, K. Janowicz, and C. A. Goble, Eds., vol. 8797. Springer, 2014, pp. 1–16.

[17] A. Isaac, H. Matthezing, L. van der Meij, S. Schlobach, S. Wang, and C. Zinn, "Putting ontology alignment in context : Usage scenarios, deployment and evaluation in a library case," in *The Semantic Web : Research and Applications, 5th European Semantic Web Conference, ESWC 2008*, ser. Lecture Notes in Computer Science, S. Bechhofer, M. Hauswirth, J. Hoffmann, and M. Koubarakis, Eds., vol. 5021. Springer, 2008, pp. 402–417.

[18] A. Isaac, D. Kramer, L. van der Meij, S. Wang, S. Schlobach, and J. Stapel, "Vocabulary matching for book indexing suggestion in linked libraries - A prototype implementation and evaluation," in *The Semantic Web - ISWC 2009, 8th International Semantic Web Conference, ISWC 2009*, ser. Lecture Notes in Computer Science, vol. 5823. Springer, 2009, pp. 843–859.

[19] D. Ritze, J. Völker, C. Meilicke, and O. Sváb-Zamazal, "Linguistic analysis for complex ontology matching," in *5th International Workshop on Ontology Matching (OM-2010)*, ser. CEUR Workshop,

P. Shvaiko, J. Euzenat, F. Giunchiglia, H. Stuckenschmidt, M. Mao, and I. F. Cruz, Eds., vol. 689, 2010.

[20] R. Parundekar, C. A. Knoblock, and J. L. Ambite, "Discovering concept coverings in ontologies of linked data sources," in *The Semantic Web - ISWC 2012 - 11th International Semantic Web Conference*, ser. Lecture Notes in Computer Science, vol. 7649. Springer, 2012, pp. 427–443.

[21] Y. An, X. Hu, and I. Song, "Learning to discover complex mappings from web forms to ontologies," in *21st ACM International Conference on Information and Knowledge Management, CIKM'12*, X. Chen, G. Lebanon, H. Wang, and M. J. Zaki, Eds. ACM, 2012, pp. 1253–1262.

[22] É. Thiéblin, F. Amarger, N. Hernandez, C. Roussey, and C. T. dos Santos, "Cross-querying LOD datasets using complex alignments : An application to agronomic taxa," in *Metadata and Semantic Research - 11th International Conference, MTSR*, ser. Communications in Computer and Information Science, E. Garoufallou, S. Virkus, R. Siatri, and D. Koutsomiha, Eds., vol. 755. Springer, 2017, pp. 25–37.

[23] L. Zhou, M. Cheatham, A. Krisnadhi, and P. Hitzler, "A complex alignment benchmark : Geolink dataset," in *The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference,*, ser. Lecture Notes in Computer Science, D. Vrandecic, K. Bontcheva, M. C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L. Kaffee, and E. Simperl, Eds., vol. 11137. Springer, 2018, pp. 273–288.

[24] A. Algergawy, M. Cheatham, D. Faria, A. Ferrara, I. Fundulaki, I. Harrow, S. Hertling, E. Jiménez-Ruiz, N. Karam, A. Khiat, P. Lambrix, H. Li, S. Montanelli, H. Paulheim, C. Pesquita, T. Saveta, D. Schmidt, P. Shvaiko, A. Splendiani, É. Thiéblin, C. Trojahn, J. Vatascinová, O. Zamazal, and L. Zhou, "Results of the ontology alignment evaluation initiative 2018," in *13th International Workshop on Ontology Matching*, ser. CEUR Workshop, P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, M. Cheatham, and O. Hassanzadeh, Eds., vol. 2288, 2018, pp. 76–116.

[25] M. Ehrig and J. Euzenat, "Relaxed precision and recall for ontology matching," in *Integrating Ontologies '05, K-CAP 2005 Workshop on Integrating Ontologies*, ser. CEUR Workshop, B. Ashpole, M. Ehrig, J. Euzenat, and H. Stuckenschmidt, Eds., vol. 156, 2005.

[26] A. Solimando, E. Jiménez-Ruiz, and C. Pinkel, "Evaluating ontology alignment systems in query answering tasks," in *ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference, ISWC*, ser. CEUR Workshop, M. Horridge, M. Rospocher, and J. van Ossenbruggen, Eds., vol. 1272, 2014, pp. 301–304.

[27] L. Hollink, M. van Assem, S. Wang, A. Isaac, and G. Schreiber, "Two variations on ontology alignment evaluation : Methodological issues," in *The Semantic Web : Research and Applications, 5th European Semantic Web Conference, ESWC 2008*, ser. Lecture Notes in Computer Science, S. Bechhofer, M. Hauswirth, J. Hoffmann, and M. Koubarakis, Eds., vol. 5021. Springer, 2008, pp. 388–401.

[28] J. David, J. Euzenat, F. Scharffe, and C. Trojahn, "The alignment API 4.0," *Semantic Web*, vol. 2, no. 1, pp. 3–10, 2011.

[29] J. Euzenat, A. Polleres, and F. Scharffe, "Processing ontology alignments with SPARQL," in *Second International Conference on Complex, Intelligent and Software Intensive Systems (CISIS-2008)*, F. Xhafa and L. Barolli, Eds. IEEE Computer Society, 2008, pp. 913–917.

[30] H. Qin, D. Dou, and P. LePendu, "Discovering executable semantic mappings between ontologies," in *On the Move to Meaningful Internet Systems 2007 : CoopIS, DOA, ODBASE, GADA, and IS, OTM*, ser. Lecture Notes in Computer Science, vol. 4803. Springer, 2007, pp. 832–849.

[31] I. Horrocks, O. Kutz, and U. Sattler, "The even more irresistible SROIQ," in *Tenth International Conference on Principles of Knowledge Representation and Reasoning*, P. Doherty, J. Mylopoulos, and C. A. Welty, Eds. AAAI Press, 2006, pp. 57–67.

[32] J. David, J. Euzenat, P. Genevès, and N. Layaïda, "Evaluation of query transformations without data : Short paper," in *Companion of The Web Conference 2018 WWW*, P. Champin, F. Gandon, M. Lalmas, and P. G. Ipeirotis, Eds. ACM, 2018, pp. 1599–1602.

[33] É. Thiéblin, F. Amarger, O. Haemmerlé, N. Hernandez, and C. T. dos Santos, "Rewriting SELECT SPARQL queries from 1 :n complex correspondences," in *11th International Workshop on Ontology Matching*, ser. CEUR Workshop, vol. 1766, 2016, pp. 49–60.

[34] M. C. Suárez-Figueroa, A. Gómez-Pérez, and M. Fernández-López, "The neon methodology for ontology engineering," in *Ontology Engineering in a Networked World*, M. C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, and A. Gangemi, Eds. Springer, 2012, pp. 9–34.

[35] O. Šváb, V. Svátek, P. Berka, D. Rak, and P. Tomášek, "Ontofarm : Towards an experimental collection of parallel ontologies," in *4th International Semantic Web Conference (ISWC). Poster*, 2005.

[36] D. Faria, C. Pesquita, B. S. Balasubramani, T. Tervo, D. Carriço, R. Garrilha, F. M. Couto, and I. F. Cruz, "Results of AML participation in OAEI 2018," in *13th International Workshop on Ontology Matching*, ser. CEUR Workshop, P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, M. Cheatham, and O. Hassanzadeh, Eds., vol. 2288, 2018, pp. 125–131.

Les anti-motifs pour l'analyse de grands graphes de connaissances incohérents

Thomas de Groot[1], Joe Raad[1,2], Stefan Schlobach[1]

[1]Department of Computer Science, Vrije Universiteit Amsterdam, Pays-Bas
[2] Université Paris-Saclay, LISN, Orsay, France

{t.j.a.de.groot, k.s.schlobach}@vu.nl, joe.raad@universite-paris-saclay.fr

Résumé

Un certain nombre de graphes de connaissances (GC) sur le Web de données contiennent des déclarations contradictoires et sont donc logiquement incohérents. Des méthodes existent pour expliquer une seule contradiction en trouvant ses justifications, qui représente l'ensemble minimal d'axiomes suffisant pour la produire. Dans les grands GC, ces justifications peuvent être fréquentes et peuvent faire référence de manière redondante au même type d'erreur en modélisation. De plus, ces justifications sont par définition dépendantes du domaine, donc difficiles à interpréter ou à comparer. Cet article utilise la notion d'anti-motif (anti-pattern) pour généraliser ces justifications, et présente une approche pour détecter presque tous les anti-motifs à partir de n'importe quel GC incohérent. Des expériences sur des GC de plus de 28 milliards de faits montrent l'évolutivité de cette approche et les avantages des anti-motifs pour analyser et comparer les erreurs logiques entre différents GC. Cet article a été déjà publié à ESWC 2021.

Mots-clés

Web des données, raisonnement, incohérence

1 Introduction

De nos jours, des graphes de connaissances (GC) de milliards de faits sont régulièrement déployés par des chercheurs de divers domaines et entreprises. Étant donné que la plupart des GC sont traditionnellement construits sur une plus longue période de temps, par différents collaborateurs, ces GC sont très susceptibles de contenir des déclarations logiquement contradictoires. En conséquence, le raisonnement sur ces GC devient limité et la connaissance formellement inutile. En générale, une fois que ces déclarations contradictoires dans un GC sont récupérées, elles sont soit expliquées logiquement et réparées, soit ignorées via un raisonnement non standard. Ce travail s'inscrit dans la première catégorie d'approches où l'accent est mis sur la recherche et l'explication de ce qui a été énoncé dans le GC qui cause l'incohérence logique. Comprendre comment ces contradictions se forment et à quelle fréquence elles peuvent se produire est essentiel pour résoudre et éviter de telles contradictions. Au moins, c'est une étape nécessaire pour développer de meilleurs outils capables de gérer des GC incohérents. Pour expliquer les contradictions, la notion de *justification*, qui est un sous-ensemble minimal du GC suffisant pour que la contradiction tienne, joue un rôle clé. Des méthodes existent pour expliquer une seule contradiction en trouvant ses justifications. Bien que les justifications fournissent une bonne base pour expliquer les problèmes de qualité des données et de modélisation dans les GC, leur spécificité dans l'explication des contradictions augmente dans certains cas la complexité d'analyse et de gestion des contradictions détectées. Particulièrement dans les grands GC, ces complexités sont amplifiées et rencontrées dans différentes dimensions :

1. Passage à l'échelle des outils. Les méthodes existantes pour récupérer les justifications de contradiction ne s'adaptent pas aux GC contenant de milliards de faits.

2. Fréquence des justifications. Les contradictions détectées avec leurs justifications peuvent être trop fréquentes pour analyser et comprendre manuellement les erreurs de modélisation commises. Ceci est particulièrement problématique lorsqu'un nombre important de ces justifications récupérées se réfèrent en fait au même type d'erreur, mais instanciées dans différentes parties du GC.

3. Dépendance au domaine des justifications. Vu que les justifications représentent un sous-ensemble du GC, elles sont par définition dépendantes du domaine et nécessitent une certaine connaissance du domaine pour comprendre la contradiction. En plus, ce fait rend la comparaison des contradictions entre différents GC plus difficiles.

Ces différents défis pour trouver et comprendre les justifications dans leur forme traditionnelle, posent les questions de recherche suivantes :

Q1 : Peut-on définir une explication plus générale des contradictions qui catégorise les erreurs les plus courantes dans un GC, indépendamment de son domaine ?

Q2 : Peut-on retrouver ces explications généralisées à partir de n'importe quel GC, indépendamment de sa taille ?

Q3 : Comment ces explications généralisées peuvent-elles aider à analyser et comparer certaines caractéristiques entre les GC les plus couramment utilisés sur le Web ?

2 Approche

Cet article publié à ESWC 2021 présente une méthode pour extraire et généraliser les justifications de tout GC incohérent. Nous appelons ces justifications généralisées des *anti-motifs* (anti-pattern) car elles peuvent être considérées comme un type d'erreurs courantes, produites soit dans les phases de modélisation ou de population du GC, soit éventuellement issues de liage erroné de données.

2.1 Exemples d'anti-motifs

Par exemple, dans l'ontologie de Pizza [1] qui sert de tutoriel pour OWL et l'éditeur d'ontologie Protégé, nous pouvons trouver deux contradictions ajoutées en exprès par ses développeurs. La contradiction (A) démontre la classe insatisfaisable *CheesyVegetableTopping*, qui a deux parents disjoints *CheeseTopping* et *VegetableTopping*. La deuxième contradiction (B) démontre une erreur courante commise, où la classe *Pizza* est affirmée comme le domaine de *hasTopping* malgré la présente d'une restriction de propriété sur la classe *IceCream* stipulant que tous les membres de cette classe doivent utiliser la propriété *hasTopping*. Cependant, comme il est également spécifié que les classes *Pizza* et *IceCream* sont disjointes, forcer maintenant une classe insatisfiable à avoir un membre conduit à une incohérence dans l'ontologie.

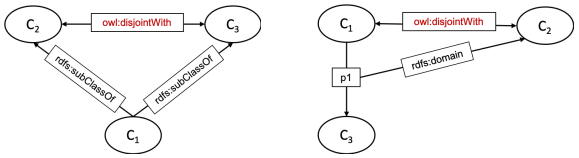
Bien que cet exemple se réfère à un cas spécifique d'une contradiction découlant de la description de ces classes, il se réfère également à un type commun d'erreur qui peut être présent dans un autre GC. Cette formalisation de certains types d'erreurs est ce que nous appelons des anti-motifs. Figure 1 présentent les deux anti-motifs généralisant les justifications des contradictions A et B. Afin de transformer une justification en anti-motif, nous remplaçons les éléments en position de sujet et d'objet du BGP par des variables (C_1, C_2, C_3 et p_1 dans figure 1). Afin d'éviter d'éliminer la contradiction, les éléments apparaissant en position prédicat d'une justification ne sont pas remplacés dans l'anti-motif, à l'exception d'un cas : les éléments apparaissant en position prédicat et apparaissant également en position sujet ou objet de la même justification.

2.2 Détection d'anti-motifs

Nous avons développé une méthode qui peut récupérer ces anti-motifs à partir de n'importe quel GC (incohérent). Garantir la récupération de *tous* les anti-motifs nécessite d'abord de trouver *toutes* les contradictions avec leurs justifications, puis de généraliser ces justifications en anti-motifs. En pratique, et comme moyen de relever les défis du passage à l'échelle, notre approche introduit un certain nombre d'heuristiques qui ne permettent pas de garantir son exhaustivité en ce qui concerne la détection de tous les anti-motifs. Cette approche est composée des trois étapes :

Étape 1. La première étape de l'approche consiste à partitionner le GC original en sous-graphes plus petits et qui se chevauchent. Selon la stratégie de partition, cette étape peut avoir un impact sur le nombre de justifications récupérées, qui à son tour peut potentiellement impacter le nombre d'anti-motifs récupérés. Nos expérimentations de la stratégie de partition sur deux GC : LOV (888,017 faits) et YAGO (158 millions faits) montre que partitionner le GC original en des sous-graphes de 5,000 triples impacte le moins le nombre de justifications récupérées.

1. https://protege.stanford.edu/ontologies/pizza/pizza.owl



(a) Anti-motif de la contradiction A

(b) Anti-motif de la contradiction B

FIGURE 1 – Représentation graphique des anti-motif des deux contradictions trouvées dans l'ontologie Pizza

Étape 2. La deuxième étape de l'approche consiste à exécuter un algorithme de récupération de justification pour détecter les contradictions avec leurs justifications. Pour cela, nous utilisons l'algorithme de récupération de justification dans le raisonneur Openllet avec le profil OWL 2 EL, qui parcourt le graphe et trouve la justification minimale pour chaque contradiction. Cette étape est exécutée pour chaque sous-graphe, et toutes les justifications sont ensuite poussées à l'étape finale du pipeline.

Étape 3. La dernière étape consiste à généraliser les justifications en anti-motifs. Les justifications ayant le "même type" sont regroupées. Par conséquent, étant donné une justification et sa généralisation en anti-motif, nous vérifions si un anti-motif avec la même structure existe déjà. Comparer des anti-motifs avec des noms de variables différents consiste à vérifier si ces anti-motifs sont isomorphes. Pour cela, nous implémentons une version de l'algorithme VF2.

Pour évaluer notre approche, nous avons montré sur des GC relativement petits que notre approche peut détecter en pratique tous les anti-motifs malgré le partitionnement du graphe, et montré sur des GC de milliards de faits que notre approche peut être appliquée à l'échelle du Web. Plus précisément, nous avons montré sur les jeux de données *LOD-a-lot* (28,3 milliards faits), *DBpedia* (1 milliards faits), *YAGO* (158 millions faits) que des milliards de justifications peuvent être généralisées en des centaines d'anti-motifs. Bien que ces résultats prouvent la propagation de milliards de faits logiquement contradictoires dans le Web des données, ce travail montre également qu'en utilisant les anti-motifs, ces contradictions peuvent désormais être facilement localisées dans d'autres GC (par exemple, en utilisant une requête *SELECT*), et éventuellement réparées (par exemple, en utilisant une requête *CONSTRUCT*). Enfin, le code source [2] de cette méthode implémenté en JAVA, ainsi que la liste des anti-motifs détectés à partir de ces GC sont en libre d'accès sous forme de requêtes SPARQL.

2. https ://github.com/thomasdegroot18/kbgenerator

Session 3 : Résultats d'études et d'états de l'art en ingénierie des connaissances

Plaidoyer pour des ontologies épistémiques

G. Kassel

Laboratoire MIS, Université de Picardie Jules Verne
33 rue Saint-Leu, 80039 Amiens Cedex 1

Gilles.kassel@u-picardie.fr

Résumé

Dans cet article, nous défendons l'usage en Ingénierie des Connaissances d'ontologies épistémiques, c'est-à-dire de systèmes de catégories représentant nos représentations du monde, plus exactement des objets de connaissance du monde, plutôt que le monde directement. En première partie, nous exposons un cadre ontologique reposant sur un double réalisme psychique (mental) et physique permettant d'étayer le développement de tels artefacts. En seconde partie, nous présentons plusieurs avantages découlant de ce cadre ontologique pour rendre compte des notions d'artefact technique, d'événement, d'action et finalement de vérité.

Mots-clés

Ontologie fondatrice, abstraits vs concrets, objets mentaux de connaissance, vérité

Abstract

In this article, we defend the use in Knowledge Engineering of epistemic ontologies, that is to say of systems of categories representing our representations of the world, more exactly world knowledge objects, rather than the world directly. In the first part, we expose an ontological framework based on a double psychic (mental) and physical realism allowing to underpin the development of such artefacts. In the second part, we present several advantages arising from this ontological framework to account for the notions of technical artefact, event, action and finally truth.

Keywords

Foundational ontology, abstract vs concretes, mental objects of knowledge, truth

1 Introduction

La question de savoir ce que représentent, ou devraient représenter, les catégories des ontologies que nous développons en Ingénierie des Connaissances a fait l'objet d'un débat en 2010 dans le journal *Applied Ontology* entre Gary Merrill [20], tenant d'une approche « conceptualiste » (les catégories représentent des concepts), et Barry Smith et Werner Ceusters [30], tenants d'une approche « réaliste » (les catégories représentent des universaux « réels » du monde physique), chacun campant de fait sur ses positions.

Ce débat se trouve être le reflet de différentes théories en philosophie du langage et de l'esprit portant sur les notions d'*intentionnalité*, de *représentation* (*connaissance*) du monde, et de *vérité* irriguant des programmes de recherche contemporains distincts [4]. Si ces théories puisent leurs racines dans l'antiquité et ont donné lieu à d'âpres discussions parmi les scolastiques du Moyen Âge [23], elles reposent sur des positions et des lignes de fractures qui ont été clairement exposées au tournant du XXe siècle, notamment dans l'école de Franz Brentano. Ses principaux disciples immédiats – Kazimierz Twardowski, Alexius Meinong et Edmund Husserl – ont en effet proposé sur un plan métaphysique des théories de l'objet en compétition [29]. Dans ces théories, la question de conférer à un objet intentionnel immanent (mental) un vrai statut ontologique et celle de reconnaître à cet objet un mode d'existence '*être pensé*' distinct de l'existence effective concrète sont centrales.

Vis-à-vis de ces questions, nous avons récemment défendu une théorie de l'*in-existence* (existence dans l'esprit) de l'objet intentionnel dans la lignée de Brentano et Twardowski, ce qui nous a conduit, dans le débat précité, à opter pour une approche conceptualiste [16, 18]. Celle-ci repose sur une thèse ontologique (TO) et une thèse sémantique (TS) d'où découle un principe méthodologique pour le développement d'ontologies (PM) :

(TO) – Des objets singuliers et généraux de pensée existent, qui permettent à des sujets conscients de se représenter et d'avoir connaissance d'aspects resp. singuliers et généraux du monde.

(TS) – les termes réfèrent indirectement à des entités réelles concrètes ; la référence est médiée par des objets immanents à la pensée.

(PM) – Les catégories ontologiques correspondent à des objets généraux de pensée ; ces objets généraux sont construits par les experts des domaines concernés ; le rôle des ontologues est de les recenser et de les organiser au moyen de liens de subsomption.

Dans cet article, nous rappelons tout d'abord nos engagements ontologiques en défense de (TO) et (TS) (§1), puis nous présentons plusieurs avantages (par rapport à une approche réaliste) découlant de ces engagements. Ainsi, nous envisageons le traitement des artefacts techniques (§2), des événements (§3) et plus spécifiquement des actions (§4). Nous développons par la suite une théorie de la connaissance et de la vérité que nous proposons comme guide pour l'ingénierie des connaissances (§5).

2 Nos engagements ontologiques

2.1 Par où débuter ?

Pour la conception d'ontologies fondatrices, et le choix des principes les plus généraux pour classer les entités du monde, deux stratégies sont couramment à l'œuvre. En Ontologie Appliquée (ingénierie ontologique), la priorité est habituellement donnée au réel physique. Cette priorité puise ses racines dans la tradition aristotélicienne de la substance et de ses accidents, ce qui conduit à distinguer entre *continuants* et *occurrents*. Dans les ontologies BFO [11] et DOLCE [19] des théories resp. *endurantiste* et *perdurandiste* sont proposées pour rendre compte de la manière distincte dont continuants et occurrents persistent dans le temps. En Métaphysique (ontologie philosophique)[1], la première distinction communément retenue est celle entre *abstraits* et *concrets* (à titre d'exemple, *cf.* en Fig. 1 l'ontologie fondatrice de Gary Rosenkrantz et Joshua Hoffman [26]). Il s'agit là de distinguer deux modes d'existence : les concrets, au contraire des abstraits, existent indépendamment de toute pensée humaine.

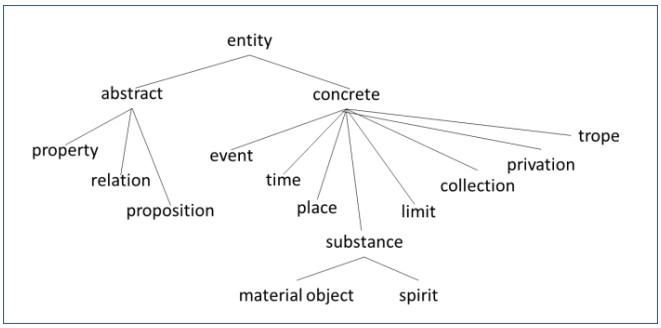


Fig. 1 : ontologie fondatrice, tiré de [26, p. 839]

Face à ces deux stratégies, nous optons pour l'option métaphysique. Mais la première est également d'inspiration métaphysique ! C'est exact, simplement sous l'influence notamment de l'école brentanienne et de la phénoménologie qu'elle a enfantée, le domaine de l'ontologie s'est élargi pour conférer une place aux contenus des états mentaux que nous éprouvons lors de nos interactions avec le monde. Ainsi, par exemple, le rouge de l'objet physique que nous percevons est-t-il distinct de la qualité physique inhérente à cet objet. L'ontologie de Rosenkrantz et Hoffman, avec la catégorie 'abstract', ménage une place pour de telles entités mentales[2]. Les engagements ontologiques que nous prenons visent précisément à clarifier la nature de cette catégorie. À cette fin, nous remontons aux travaux de Brentano et Twardowski.

2.2 La doctrine intentionnaliste de Brentano

La doctrine de l'intentionnalité de Brentano (tout du moins celle précédant sa phase dite *réiste*, dont nous nous revendiquons ici), nous est connue principalement par son essai [7] *Psychologie du point de vue empirique* et par l'ouvrage [8] *Brentano. Deskriptive Psychologie* édité par Roderick Chisholm et Wilhelm Baumgartner rassemblant des travaux menés par Brentano dans les années 1880 et 90. Visant à fonder le domaine de la psychologie, Brentano caractérise les phénomènes (ou actes) de pensée comme « se référant à un contenu » ou « étant dirigés vers un objet », chaque type d'acte (présentation, jugement, mouvement de l'âme) ayant son propre type de contenu. Nous notons sa théorie ontologique de l'intentionnalité comme un modèle à 4 termes : acte / contenu mental / objet mental / chose extra-mentale[3].

Pour illustrer ce modèle, considérons le fait, pour un sujet, de penser à un objet matériel physique *A*. Pour le Brentano de 1874, une telle pensée consiste à se *présenter* (*vorstellen*) l'objet *A* au moyen d'un « contenu-objet » mental (nous plaçons des guillemets pour indiquer qu'en 1874, Brentano ne distingue pas encore le contenu de l'objet). La nature représentative de cette entité ne fait par contre aucun doute[4]. Cognitivement, une des fonctions de ce contenu-objet est de permettre de penser à un *A* sans que celui-ci soit concomitamment existant, qu'il soit passé ou un possible futur. Arkadiusz Chrudzimski [9] évoque à ce propos un cueilleur d'un champignon vénéneux qui, l'ayant placé dans un réfrigérateur, continue à penser à ce champignon alors qu'à son insu une personne de son entourage l'aura sorti du réfrigérateur et détruit pour s'en débarrasser.

Pour caractériser le contenu-objet, Brentano fait appel à la notion médiévale d'*objet*. L'*objet* est l'objet d'une activité intellectuelle, laquelle consiste à saisir une chose (un objet réel) selon un certain aspect (du reste, une même chose peut être pensée selon plusieurs aspects, donnant lieu à des objets de pensée distincts). L'objet est *formel*, au sens aristotélicien. Selon la psychologie d'Aristote (présentée dans *De Anima*), la perception d'un objet réel revient en une nouvelle actualisation de la forme de l'objet, dissociée de la matière, dans le sujet. Brentano aménage le principe sur deux points : d'une part, les formes perçues ne sont pas portées par le sujet lui-même mais par un corrélat intentionnel de la chose *A* (un *ens rationis*, pour

[1] Pour notre propos dans cet article, nous utiliserons les termes « métaphysique » et « ontologie » comme synonymes.

[2] Tout en considérant que, selon ces auteurs, les fondements métaphysiques de la distinction entre *concrets* et *abstraits* n'est pas bien définie [26] : "the distinction between concrete entities and abstract entities that we are employing is an intuitive one. The intuitive distinction may be difficult to analyze, but it is serviceable nonetheless. The distinction in question seems indispensable in ontology, and is presupposed by realists and antirealists in their debates about the problem of universals".

[3] De nombreux commentateurs de Brentano ne retiennent qu'un modèle à 3 termes, excluant l'objet mental (*cf.*, concernant les diverses interprétations des écrits de Brentano, les éclaircissements apportés par Guillaume Fréchette [10]). Dans cet article, nous nous fondons principalement sur l'analyse de Mauro Antonelli [2], lequel défend un modèle à 4 termes. Selon Antonelli, ce n'est que dans la phase dite *réiste* clôturant sa production scientifique que Brentano renoncera à conférer au « corrélat intentionnel » de l'acte un statut ontologique.

[4] La plupart des commentateurs et traducteurs de Brentano ne s'y tromperont pas en utilisant le terme « représentation », là où l'étymologie du terme allemand *vorstelle*n – placer (*stellen*) devant (*vor*) – plaiderait plutôt en faveur du terme « présentation ». De fait, il faut considérer que cette entité se *présente* à la conscience (perception interne) et remplit en même temps une fonction de *représentation* vis-à-vis d'une autre entité transcendante à l'acte.

les scolastiques du Moyen Âge) ; d'autre part, les déterminations de ce corrélat ne coïncident pas avec la forme des choses, elles en sont une représentation. Dans les années 1880, Brentano propose une analyse méréologique de ce corrélat intentionnel, noté dorénavant *A pensé*. La notation traduit le fait que l'objet *A réel* est présent en tant qu'objet modifié dans le corrélat et qu'il reste ainsi présent à la conscience du sujet[5].

Évoquons maintenant le *jugement* (*Urteil*). Selon Brentano, le jugement de base (auquel se ramènent les jugements plus complexes) s'exprime sous la forme « *A* existe », à l'instar du jugement emblématique « Dieu existe ». Brentano rompt avec la doctrine traditionnelle (au moins aristotélicienne) du jugement comme association d'un sujet et d'un prédicat et considère le jugement comme le lieu d'acceptation ou de rejet de son objet. Le terme « existe » de l'expression susmentionnée ne correspond dès lors pas à un prédicat attribué à un objet. La notion d'existence convoquée ici se comprend dans le cadre d'une théorie de la vérité fondée sur un principe de correspondance entre une chose externe et un *A pensé* (*adaequatio rei et intellectus*). Un jugement positif tel « *A* existe » revient à admettre ou reconnaître (*anerkennen*) *A*, autrement dit à considérer qu'il correspond à l'objet *A* une chose jouissant d'une réalité effective. A contrario, un jugement négatif tel « *A* n'existe pas » revient à rejeter ou renier (*verwerfen*) *A*, autrement dit à considérer qu'il ne correspond pas à l'objet *A* une chose concrète. L'objet du jugement demeure celui de la représentation, seule la modalité intentionnelle vis-à-vis de cet objet change. Dans le cas d'un jugement catégorique s'exprimant par « *A* est *B* », tel « l'arbre est vert », nous avons affaire à un jugement multiple : reconnaître *A*, reconnaître la propriété *B* (par exemple, identifier une qualité réelle effective) et reconnaître un lien unissant ces entités effectives. Pour Brentano, une spécificité du jugement, qui distingue cet acte de la présentation, est sa polarité liée à l'opposition entre acceptation et rejet d'un objet. À cette polarité s'ajoute le caractère de vérité 'vrai' ou 'faux' du jugement, une qualité du jugement dépendant des preuves mobilisées pour accepter ou renier l'objet.

Nous venons de décrire dans les grandes lignes la théorie ontologique brentanienne concernant, d'une part, l'acte consistant pour un sujet à penser à un objet transcendant son esprit et, d'autre part, l'acte consistant à juger de l'existence de l'objet. Nous présentons dans la section suivante les extensions apportées par Twardowski à ces théories.

2.3 L'objet de la représentation de Twardowski

Venons-en à Twardowski, dont la théorie de l'intentionnalité – dans la lignée de celle de Brentano – nous est connue principalement par sa thèse d'habilitation [33] *Sur la théorie du contenu et de l'objet des représentations*. Sa motivation première est de rendre compte des représentations *anobjectuelles* (ne possédant pas de référence effective) bolzaniennes, ce qui le mènera à adopter une position originale au sein de l'école brentanienne quant à l'ontologie de la représentation. Amorcée également dans ce texte, on trouve une extension de l'objet du jugement avec la prise en compte de complexes ou d'états d'affaires.

Parmi lesdites représentations anobjectuelles figurent des représentations ne référant à aucune entité rencontrée jusqu'à présent, comme [la montagne d'or], et des représentations comportant des déterminations contradictoires, comme [le carré rond]. Le geste décisif de Twardowski est de considérer que ces représentations ont un contenu se rapportant à un objet « n'existant pas ». L'argument est le suivant : lorsque nous pensons à une montagne d'or ou à un carré rond, nous pensons bien à quelque chose et ce quelque chose ne peut être le contenu de la représentation (ce n'est pas de ce dernier dont on pense que c'est une montagne et qu'il est constitué d'or) ; c'est donc bien aux objets 'montagne d'or' et 'carré rond' auxquels nous pensons et auxquels il convient de conférer un statut ontologique d'objet pensé. Ce sont les *A* des *A pensés* (selon le modèle brentanien). Dès lors, l'expression « n'existant pas » est à entendre dans le sens d'une visée judicative de l'objet (selon la doctrine brentanienne du jugement existentiel) revenant à dénier l'objet : il ne correspond à ces objets aucune chose effective. En résumé, ces représentations ont bien un objet immanent, du domaine du quelque chose (Twardowski en profite pour faire entrer dans ce quelque chose les objets impossibles), même si certaines ne réfèrent pas. Ce geste de Twardowski est capital dans la mesure où il consacre une ontologie comportant deux modes d'être (d'existence) distincts, l'*être pensé* et l'*être effectif*. L'être pensé est l'existence conférée par toute représentation à son objet, y compris donc pour les représentations ne référant pas extérieurement à des choses effectives. Ce même mode d'existence s'applique à la représentation globalement – le *A pensé* – porteuse de propriétés spécifiques comme celle d'existence effective (attribuée lors d'un jugement). On peut donc parler d'existence mentale, par opposition à l'existence effective. Le modèle à 4 termes – acte / contenu mental / objet mental / chose (externe) – attribué par Antonelli à Brentano, est confirmé et généralisé.

L'essai de Twardowski de 1894 consacre par ailleurs une part importante à l'étude de la structure méréologique des objets pensés au moyen de relations qualifiées de formelles et de matérielles, Twardowski s'intéressant de la sorte à des objets complexes ou *états de choses* (*Sachverhalt*)[6]. L'intérêt porté par Twardowski à ces états de choses tient au rôle qu'il leur attribue comme objets de jugements relationnels exprimés par des phrases comme « la table est blanche » ou « Paul déplace la table ». L'idée est que ces jugements ont pour objet principal, respectivement 'l'être blanc de la table' et 'le déplacement de la table par Paul'. Le cadre d'analyse demeure celui de la doctrine brentanienne du jugement consistant à accepter ou

[5] Selon Antonelli [2, p. 40] : "(…) the object, unlike the correlate, is not only part of the mental act in a *modified sense* but also, contemporaneously, part of the *correlate itself*, again in a modifying sense". En d'autres termes [2, p. 34) : "(…) In inner perception the intentional object (the primary object) does not fade out of the consciousness; it is still there, but embedded in a more complex structure of which it is only a moment or a part".

[6] Arianna Betti, dans son [5] *Propositions et états de choses chez Twardowski*, nous indique que Twardowski a complété sa théorie de l'état de choses à l'occasion d'un cours de logique qu'il a donné à Vienne à l'hiver 1894-1895 (dont les notes ont été préservées et traduites récemment en allemand par Betti et Venanzio Raspa [6]).

rejeter son objet. Les jugements sont analysés comme (resp.) « l'être blanc de la table existe » et « le déplacement par Paul de la table existe », l'existence revenant à vérifier si l'état de choses pensé correspond bien à un état de choses effectif[7]. Il est important de noter que ces états de choses s'avèrent distincts de relations, mais également de propositions car ils ne portent pas de valeur de vérité (et, de fait, les propositions [la table est blanche] et [la table n'est pas blanche] ont le même état de choses comme objet principal).

2.4 Nos engagements, en bref

Brentano et Twardowski, rappelons-le, ont évolué au cours de leur vie de chercheur dans les thèses psychologiques et ontologiques qu'ils ont défendues (avec, du reste, plus ou moins de réussite) et d'autres disciples de Brentano ont défendu des thèses alternatives pour rendre compte du phénomène de l'intentionnalité[8]. Dans les sections §2.2 et §2.3, nous avons fait en sorte de reprendre les thèses que nous souhaitons nous approprier et que nous résumons maintenant (en les illustrant en Fig. 2).

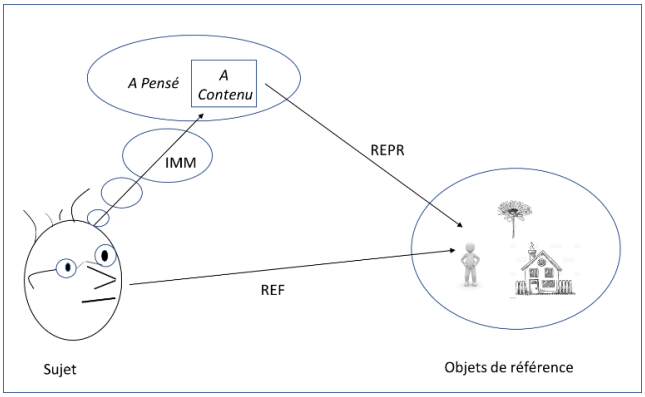


Fig. 2 : notre ontologie de l'acte de pensée à quelque chose d'extra-mental

Pour penser le monde et le connaître, un sujet entretient des entités mentales ressortissant de plusieurs catégories : des propriétés/relations (non présentées dans la figure) et des objets particuliers et généraux. Ces différents types d'entités sont liées : les propriétés sont attachées aux objets en constituant leur « être-tel » ; un objet est ainsi conçu comme ayant telle et telle propriété et/ou en étant en relation avec tel et tel objet ; *l'in-existence* de l'objet ne présume pas l'existence d'objets extra-mentaux avec lesquels il peut entretenir une relation de représentation (principe d'indépendance de l'être-tel vis-à-vis de l'existence effective d'objets représentés). Lorsque le sujet pense ou se réfère à ce qu'il conçoit être une chose extra-mentale (rel *REF*), sa conscience se dirige vers l'objet de connaissance immanent (rel *IMM*) représentant cet objet (rel *REPR*). La relation *IMM* tient pour une intentionnalité directe de l'acte de pensée : au cours de l'acte, l'objet est directement présenté à la conscience (perception interne). En revanche, la relation *REF* tient pour une intentionnalité indirecte. Dans le cas où la chose extra-mentale n'existe pas, la visée s'arrête à l'objet immanent.

Dans la suite de l'article, nous dégageons plusieurs avantages que nous attribuons à cette théorie de l'intentionnalité indirecte.

3 Les artefacts techniques comme objets de représentation

À titre de première illustration de notre cadre ontologique et de son apport, notamment par rapport à l'approche « réaliste » de BFO, considérons le domaine des artefacts, communément conçus comme comportant une dimension sociale. La philosophe Amie Thomasson [33] distingue deux principales catégories d'entités sociales. D'une part, des entités sociales *concrètes*, à savoir des entités concrètes sur lesquelles surviennent des faits sociaux, ce qui est le cas pour les artefacts techniques matériels (ex : une bouteille, un presse-papier, un tournevis) : ces entités jouent le rôle du *Y* dans la règle constitutive searlienne [28] « *X* compte pour *Y* dans le contexte *C* ». D'autre part, des entités sociales *abstraites* (ex : une loi, une monnaie, un syndicat) pour lesquelles la règle searlienne ne s'applique pas, faute de pouvoir exhiber un *X* sur lequel surviendrait un fait social.

Focalisons-nous sur les artefacts techniques en prenant comme exemples un bistouri et un scalpel[9]. La question posée est de décider quelles entités dans les domaines mentaux et concrets peuvent en rendre compte. Il se trouve que ces deux artefacts sont des objets *physiques* strictement identiques. Ils jouent dès lors le rôle d'un même *X* pouvant compter pour un BISTOURI lors d'une opération ou pour un SCALPEL lors d'une dissection. Toujours selon Searle, « compter pour » dans ces cas revient à leur attribuer une fonction, à savoir une propriété « relative à un observateur » (à distinguer d'une propriété physique « inhérente »). Il n'est donc pas possible, sauf à renier la dimension sociale des artefacts, de les faire correspondre (uniquement) à des objets physiques[10]. Nous considérons dès lors deux entités : d'une part, côté concret, l'objet physique (le *X*) *simpliciter* ; d'autre part, côté mental/social, l'artefact pensé comme étant un objet auquel sont attribuées des propriétés

[7] Pour être complet, mentionnons le fait que Twardowski a également une théorie de l'objet général ou « universel ». L'objet de la représentation [homme] est analogue à une idée platonicienne, à ceci près qu'elle est mentale. Cet objet se réfère à des hommes singuliers concrets par l'intermédiaire de représentations singulières comme [Paul]. Cette généralité, appliquée à des états de choses, conduit à considérer des états de choses universaux objets de représentations comme [l'être mûr d'un fruit] ou [l'amour porté par une personne à une autre personne].

[8] Dans [18], nous développons des argumentaires pour rejeter lesdites alternatives, notamment la théorie de l'objet de Meinong et l'ontologie de Husserl.

[9] Nous empruntons à Bruno Bachimont et Jean Charlet ce couple d'exemples (fréquemment cité dans leurs écrits) compte tenu de sa particularité mentionnée dans la suite du paragraphe.

[10] C'est pourtant ce que proposent les auteurs de BFO lorsqu'ils évoquent les références retenues pour définir leur notion de *fonction* [30, note 4) : « À noter que toutes correspondent à des vues réalistes : elles considèrent que les fonctions existent, qu'elles sont des ingrédients de l'être. Nous ne tenons pas compte de ces traitements – défendus par exemple par Searle (1995) – selon lesquels le discours fonctionnel est une *simple façon de parler* à propos des choses et ainsi en principe éliminables ».

physiques et sociales. De telles attributions correspondent à des stipulations mentales et sont à distinguer de l'inhérence de propriétés.

4 Les événements comme objets de représentation

Venons-en aux événements, auxquels nous reconnaissons une nature mentale / sociale [14][15]. Plus précisément, nous distinguons les événements mentaux de faits dont l'existence peut être assimilée (mentalement) à la survenue de ces événements.

La thèse défendue ici est d'identifier les événements aux états de choses mentaux caractérisés par Twardowski (*cf.* §2.3). Dès lors nous pouvons mettre en avant une notion d'*occurrence* associée à celle d'*existence* telle que définie dans le cadre du jugement existentiel brentanien : juger de l'occurrence d'un événement revient à reconnaître (ou à rejeter) que des faits correspondant aux conditions de satisfaction de l'événement tiennent (existent).

À titre d'illustration, reprenons les exemples d'événements donnés en §2.3, à savoir 'l'être blanc de la table' et 'le déplacement de la table par Paul' dont les phrases (1a) et (2a) ci-dessous expriment l'occurrence. Pour identifier les faits « occurrent-facteurs » en (1b), nous faisons appel à la théorie des tropes en conférant une existence physique à la qualité individuelle '$\text{Blanc}_{\text{Table}}$'. Le fait que la même qualité inhère à la table à différents instants traduit le fait que l'événement soit un état. L'exemple (2b) est plus complexe car il fait intervenir d'une part un processus '$\text{Proc}_{\#i}$' énacté par 'Paul' (sans doute un geste corporel) et d'autre part un processus de déplacement de la table '$\text{Déplacement}_{\#j}$' engendré causalement par '$\text{Proc}_{\#i}$'[11].

(1) a « La table est blanche »

b <Inhère, Table, $\text{Blanc}_{\text{Table}}$, I_1>, <Inhère, Table, $\text{Blanc}_{\text{Table}}$, I_2>, …

(2) a « Paul déplace la table »

b [<Énacte, Paul, $\text{Proc}_{\#i}$, I_1>, <Énacte, Paul, $\text{Proc}_{\#i}$, I_2>,…] ; [<Énacte, Table, $\text{Déplacer}_{\#j}$, I_1>, <Énacte, Table, $\text{Déplacer}_{\#j}$, I2>,…] ; [<Perpétue, $\text{Proc}_{\#i}$, $\text{Déplacer}_{\#j}$, I_1>, <Perpétue, $\text{Proc}_{\#i}$, $\text{Déplacer}_{\#j}$, I_2>, …]

Outre ce traitement de l'occurrence d'événement, évoquons deux apports que nous voyons à cette conception mentale des événements, par rapport à la figure prédominante de l'événement concret.

Un problème intriguant de nombreux philosophes (une « nuisance philosophique » selon Neil Wilson [37, p. 305], une « bizarrerie » selon Peter Hacker [12, p. 14]) est que certains événements comme un pique-nique, une bataille ou un ouragan, semblent « se mouvoir ». Nos engagements ontologiques permettent d'apporter une explication à ces puzzles. Selon ces engagements, les événements sont des entités conçues par des sujets et relatent l'histoire du monde à partir de l'interprétation de faits auxquels contribuent d'autres entités. Les événements cités correspondent de fait à l'histoire d'entités concrètes changeant de place. Ce sont donc ces entités concrètes (à l'instar des participants à un pique-nique ou de la table déplacée par Paul) qui se meuvent, et non les événements.

Un autre problème auquel nous attachons de l'importance est lié au fait d'ouvrir l'inventaire ontologique à des entités 4D, c'est-à-dire à des entités dont l'essence est fondée sur le fait d'être étendues temporellement. Les événements, en métaphysique, sont justement couramment assimilés à des entités 4D selon la théorie du perdurantisme stipulant que ces entités tiennent leur existence du fait de gagner temporellement des parties. Ainsi, prenons l'exemple d'un match de football auquel nous assistons à un instant *t* : selon cette théorie, à cet instant *t*, seule une partie du match existe ; les parties antérieures n'existent plus ; les parties à venir n'existent pas encore[12]. Contre cette conception de la persistance d'un événement concret (dont on notera qu'il n'est méréologiquement complet que lorsqu'il n'existe plus du tout !), la conception mentale permet de considérer que joueurs, spectateurs et commentateurs pensent jouer ou assister à un match[13]. De fait, le match existant en pensée, ceci permet d'expliquer que les spectateurs aient pu acheter des billets avant que le match ne se déroule (occurre).

5 Les actions, en tant qu'événements intentionnels

En prolongement de la présentation des événements, disons quelques mots des actions, traditionnellement assimilées à des événements contrôlés intentionnellement. Notre conception de l'événement mental permet d'assimiler celui-ci au contenu de l'intention contrôlant l'action, traduisant le caractère autodescriptif de l'action tel que déterminé par Searle [27].

L'événement-action, par exemple le déplacement intentionnel de la table par Paul, est toutefois distinct de l'événement-contenu de l'intention contrôlant l'action. De fait, selon les travaux de philosophes de l'action, notamment ceux de Elisabeth Pacherie [22], nous sommes en présence de plusieurs intentions (*distale*, *proximale*, *motrice*) ayant des contenus et des formats différents et s'actualisant en fonction du déroulement de l'action. Seules les intentions distale et proximale ont un format conceptuel correspondant à notre notion d'événement. Considérons que Paul conçoive l'intention de déplacer le lendemain la table. Selon nos engagements et la théorie des objets généraux de pensée de Twardowski, le contenu de l'intention *distale* de Paul est alors un événement général faisant abstraction de détails comme l'heure et l'ampleur du déplacement. Au cours de l'occurrence de l'action, celle-ci est également contrôlée par une intention

[11] Ces exemples (et plusieurs autres) sont détaillés dans [17].

[12] Nous reprenons ici la théorie perdurantiste adoptée dans DOLCE [19]. À noter que dans l'ontologie BFO [11], les occurrents (les entités SPAN) sont assimilés à des entités 4D.

[13] Qui plus est au *même* match, mais il faudrait faire appel à une théorie de l'intentionnalité collective pour expliquer cette identité, ce que la place de l'article ne permet pas.

proximale ayant pour contenu un événement plus déterminé prenant en compte des informations liées à la situation en cours.

Un apport de cette conception de l'action est de pouvoir rendre compte de couplages temporels et causaux entre entités mentales et physiques (dans le cas d'une action physique) créant des boucles rétroactives [13]. D'une part, les intentions, en fonction de leurs événements-contenus, sont à l'origine de processus physiques modifiant l'état du monde. En retour, la perception par l'agent de ces modifications provoque l'actualisation des événements-contenus de ses intentions. Selon cette description, les événements jouent bien leur rôle d'apporter à l'agent une connaissance de l'évolution du monde.

6 Une théorie de la connaissance / vérité comme correspondance avec le monde

Dans cette dernière section, nous abordons les notions de *connaissance* et de *vérité*. L'objectif est d'évaluer l'adéquation du cadre ontologique posé en §2, et complété dans les sections suivantes, comme socle pour permettre d'élaborer une théorie de la connaissance et de la vérité. Pour des raisons de place, l'objectif plus modeste que nous visons est d'apporter un éclairage en nous référant aux théories de la vérité de Brentano et de Twardowski. Celles-ci ont fait l'objet d'analyses récentes respectivement par Antonelli [1] et Sébastien Richard [25].

Pour aborder la discussion, nous entendons les notions de *connaissance* et de *vérité* dans les sens (larges) suivants : la connaissance a à voir avec le fait de disposer d'une représentation du monde ; la vérité a à voir avec la conformité de ces représentations au monde. De tout temps, et au moins depuis Aristote, ces notions renvoient à une correspondance entre pensée et réel, ainsi que l'exprime Aristote dans la *Métaphysique* [3 p. 54] :

> […] être dans le vrai, c'est penser que ce qui est séparé est séparé, et que ce qui est uni est uni ; être dans le faux, c'est penser contrairement à la nature des objets.

Brentano et Twardowski sont tous deux persuadés que la vérité repose sur une forme de correspondance mais posent d'emblée une limite intrinsèque à cette notion. On le voit exprimé par Brentano dans ce fameux passage de sa *Psychologie* [7, p. 33] :

> Les phénomènes qu'il [le physicien] étudie et qui concernent la lumière, le son, la chaleur, le lieu, le mouvement local n'ont pas d'existence véritable (…). Ils constituent les signes d'une réalité effective dont l'action produit leur représentation. Mais l'image qu'ils en donnent ne correspond aucunement à cette réalité, et la connaissance qu'on en peut tirer demeure bien imparfaite. Nous pouvons dire qu'il existe quelque chose qui, dans telles ou telles conditions, devient la cause de telle ou telle sensation ; nous pouvons également démontrer qu'il doit s'y rencontrer des relations analogues à celles que représentent les manifestations spatiales, les grandeurs et les formes. Mais il faut s'en tenir là. La vérité des phénomènes physiques n'est, suivant l'expression consacrée, qu'une vérité relative.

Selon Brentano, nous n'avons qu'une connaissance imparfaite du réel nous environnant et ceci vient du fait que notre perception externe est limitée, voire faillible[14]. De ce fait, dans l'incapacité où nous nous trouvons de comparer les deux membres en correspondance, qu'exige la notion de vérité, celle-ci n'est que relative.

Pour tâcher de caractériser cette vérité relative, Brentano toute sa vie se confrontera à ce principe d'« accord » ou de « conformité » avec la réalité « *Veritas est adaequatio intellectus et rei* » [1]. Sa conception de la vérité, nous l'avons vu, s'incarne dans le jugement, qu'il définit comme un acte psychique d'acceptation (ou de refus) d'un objet de représentation, le jugement élémentaire selon lequel « *A* existe » étant vrai ssi *A* correspond à un objet réel. Le cas d'une attribution d'une propriété à un objet [*A* est *b*] tient lieu de jugement complexe composé de plusieurs jugements élémentaires. Cette conception du jugement existentiel fondée sur une correspondance est toutefois grevée de difficultés. Des difficultés concernent les jugements faux, qu'il s'agisse de jugements affirmatifs [l'arbre est bleu] ou infirmatifs [l'arbre n'est pas bleu]. Mais il est possible de considérer qu'un jugement faux revient à évaluer une absence de correspondance. D'autres difficultés concernent des jugements référant au passé ou au futur, comme avec [Paul pense à Aristote], ou référant à des irréels, par exemple [Pégase n'existe pas]. Bien qu'on puisse considérer ces derniers jugements comme vrais, ils posent un problème ontologique : comment évaluer une quelconque correspondance lorsque l'objet (ex : Pégase) n'existe pas ?

Le traitement des objets non-existants de Twardowski répond en partie au problème. Pégase existe en tant qu'objet de représentation mais il ne représente pas d'objet réel : le jugement [Pégase n'existe pas] est donc vrai. Par contre, on peut se demander à quelles entités réfèrent les jugements : [Paul pense à Aristote] et [Pégase est un cheval ailé] ? Le premier est un état psychique de Paul, le second un fait social relevant de la mythologie. Ces exemples montrent que le principe de correspondance, s'il est maintenu, doit être étendu. La raison en est simple : le monde auquel nous faisons référence, et à propos duquel nous communiquons, ne se limite pas au monde biologique et physique.

En lien avec la référence au social, nous avons vu que nos engagements ontologiques vis-à-vis des événements nous ont conduits à étendre le principe de correspondance. Plutôt qu'une relation 1:1, nous proposons une relation 1:n. Cette extension relève selon nous d'une contrainte épistémique : la connaissance de la façon dont le monde évolue (de ses stabilités comme de ses changements) nécessite qu'un événement soit mis en correspondance avec plusieurs faits tenant (existant) à des instants différents.

À ce stade de la discussion, nous avons considéré, côté pensée, comme objets porteurs de connaissance et de vérité d'une part des objets se référant à des choses concrètes et d'autre part des objets – des événements – se référant à une multitude de faits. En philosophie contemporaine du langage et de la pensée, le porteur de vérité privilégié est la *proposition*, le débat sur l'intentionnalité portant désormais presque exclusivement sur

[14] Brentano prend ici des distances avec la théorie psychologique de la perception externe d'Aristote (telle qu'exposée dans *De Anima*).

les attitudes propositionnelles[15]. Dans l'école brentanienne, le palier propositionnel sera introduit par Meinong avec son *objectif* (*Objektiv*) et l'*assomption*, à savoir l'acte de présentation de l'objectif[16]. Twardowski considère la proposition dans son [34] *Fonctions et Formations*, comme produit de la pensée, notamment du jugement. Nous l'évoquons ici car la considération par Twardowski de ce palier l'a conduit à l'hiver 1924-25 dans [35] *Theory of knowledge* à formuler une conception originale de la vérité des propositions.

Une question posée est de savoir si, en complément de la relativité de la vérité liée à notre connaissance limitée du réel (déjà évoquée), vient s'ajouter une relativité liée aux circonstances du jugement (comme avec « il pleut ») ou au sujet auteur du jugement (comme pour « l'odeur de cette fleur est plaisante »). Twardowski, contrairement à Brentano, défend sur cette question le caractère absolu de la vérité du jugement. Sa thèse repose sur deux points : d'une part, (i) l'énoncé du jugement doit être distingué du jugement lui-même, en tant que proposition vériporteuse ; d'autre part, (ii) la proposition a pour constituants les circonstances spatio-temporelles et le sujet qui en est l'auteur. Ainsi peut-on considérer qu'un sujet énonçant la phrase « il pleut » énonce en fait une proposition comme [Il pleut sur Paris, au 26 rue de la Paix à 15H15] et qu'un sujet énonçant la phrase « l'odeur de cette fleur est plaisante » énonce en fait la proposition [L'odeur de cette fleur est plaisante pour moi]. Dans les deux cas, selon (i), le sujet, par souci d'économie, n'exprime qu'une partie du jugement. Cette question, comme on le voit, concerne la nature de la proposition et le « problème de son unité » sur lequel existe une littérature abondante.

En résumé de cette discussion, précisons ce que nous retenons comme principes ontologiques pour fonder une théorie correspondantiste de la connaissance et de la vérité. Nous retenons nos objets de représentation comme modèle du monde avec tout ce que le terme « modèle » convoie comme simplifications liées à l'abstraction. Parmi ces objets figurent des représentations d'objets et de processus physiques et de leurs qualités, ces entités concrètes pouvant être actuelles, passées ou de possibles futurs. Toujours parmi ces objets figurent des événements représentant des évolutions (états et changements) du monde, actuelles, passées ou possibles. Ces objets de connaissance constituent un socle sur lequel se fondent les jugements et leurs contenus, des propositions. On notera à ce propos que la question de la nature des propositions évoquée supra se pose également pour les événements. Quoi qu'il en soit, nos propositions, à l'instar des événements, sont mentales. Établir leur vérité consiste en un raisonnement. Nous nous distinguons dès lors d'une conception classique de la vérifaction fondée sur une nécessité logique[17]. Dans l'article, nous avons envisagé une classe de propositions correspondant à des affirmations / infirmations d'occurrence d'événement singulier. Il reste d'autres classes de propositions à considérer comme celles exprimant des lois physiques (ex : « les matériaux chauffés se dilatent ») ou des lois économiques (comme : « l'inflation provoque des baisses de pouvoir d'achat »). Ceci nécessitera de compléter notre ontologie en s'engageant sur la nature des lois et en s'ouvrant à la réalité sociale [32].

7 Conclusion

En guise de conclusion, sont rassemblées en Fig. 3 les différentes entités mentionnées dans l'article. Le lecteur rapprochera l'ontologie de la Fig. 3 de celle de la Fig. 1 présentée en Introduction.

Le chemin parcouru a consisté à préciser la nature des catégories des ontologies. Il s'agit d'objets généraux de représentation, autrement dit d'objets mentaux représentant pour un sujet ses connaissances du monde, ce qu'exprime la qualification « ontologie épistémique ». Le sujet n'est toutefois pas particularisé, entraînant que l'on puisse élaborer des ontologies épistémiques rendant compte de schèmes conceptuels correspondant à des cultures diverses. Un des enjeux, déjà reconnu en Ingénierie des Connaissances, est de disposer de telles ontologies de référence qui soient partagées par différents systèmes d'information en réutilisant notamment un système de catégories abstraites, c'est-à-dire une ontologie fondatrice.

Concernant l'ontologie fondatrice présentée dans cet article, son développement est en cours et plusieurs chantiers sont ouverts, notamment celui de caractériser le domaine des entités mentales. En effet, en ne retenant que deux modes principaux d'existence, *être pensé* (mental) et *être effectif* (concret), nous avons implicitement rangé parmi les entités mentales les propriétés et relations, les entités fictives (tels les personnages de romans ou les figures mythologiques) ainsi que les entités idéales (à l'instar des entités mathématiques). C'est sur ce chantier en particulier que nous faisons porter actuellement nos efforts.

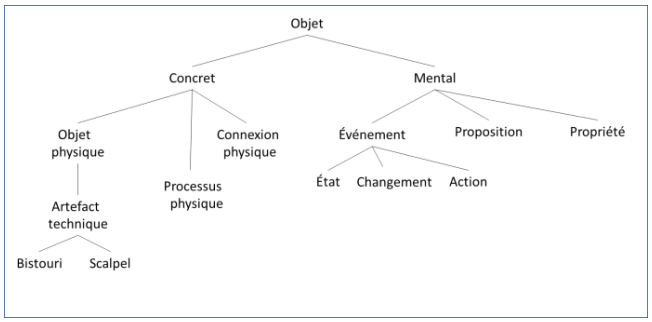


Fig. 3 : esquisse d'une ontologie épistémique

8 Références

[1] M. Antonelli, La conception de la vérité du jeune Brentano, dans J. Benoist (éd.), *Propositions et états de choses : entre être et sens*, Paris : J. Vrin, pp. 67-86, 2006.

[2] M. Antonelli, Franz Brentano's Intentionality Thesis. A New Objection to the "Nonsense that was Dreamt up and

[15] On le constate aussi bien chez Russell et Wittgenstein que chez les brentaniens comme Meinong et Twardowski [24]. La raison est que « juger que *p* » nécessite de comprendre *p* et donc d'en avoir une représentation avant de lui attribuer une valeur de vérité.

[16] Chez Meinong, l'objectif est toutefois une entité extra-mentale pouvant être saisie par la pensée.

[17] Comme on peut le voir notamment dans [21].

Attributed to him", *Brentano Studien*, Vol. 13, pp. 23-53, 2015.

[3] Aristote, *Métaphysique. Tome 2 – livres H-N*, trad. fr. et notes J. Tricot, Paris : J. Vrin, 1991.

[4] J. Benoist (dir.), *Propositions et états de choses. Entre être et sens*, Paris, Librairie Philosophique J. Vrin, 2006.

[5] A. Betti, Propositions et états de choses chez Twardowski, *Dialogue*, Vol. 14, pp. 469-92, 2005.

[6] A. Betti et V. Raspa, Kazimierz Twardowski. Logik: Wiener Logikkolleg, 1894-95, *Phenomenology & Mind*, Vol. 17, 2016.

[7] F. Brentano, *Psychologie du point de vue empirique*, Aubier, Paris, 1944 ; 2° éd revue par J. Fr. Courtine, Vrin, Paris, 2008 ; trad. fr. par M. de Gandillac de *Psychologie vom empirischen Standpunkt*, vol. I, O. Kraus (ed.), Leipzig: Meiner, 1874.

[8] R.M. Chisholm et W. Baumgartner (eds.), *Brentano. Deskriptive Psychologie*, Hamburg: Meiner, 1982.

[9] A. Chrudzimski, Brentano and Aristotle on the Ontology of Intentionality, dans D. Fisette et G. Fréchette (eds.), *Themes from Brentano*, Amsterdam : Rodopi, pp. 121-137, 2013.

[10] G. Fréchette, Brentano's Conception of Intentionality, New facts and Unsettled Issues, *Brentano Studien*, Vol. 13, pp. 9-21, 2015.

[11] P. Grenon et B. Smith, SNAP and SPAN: Towards dynamic spatial ontology, *Spatial Cognition and Computation*, Vol. 87, pp. 69-103, 2004.

[12] P.M.S. Hacker, Events and Objects in Space and Time, *Mind*, Vol. 91, N° 361, pp. 1-19, 1982.

[13] G. Kassel, Processus, événements et couplages temporels et causaux, *Revue d'Intelligence Artificielle*, Vol. 31, N° 6, pp. 649-679, 2017.

[14] G. Kassel, Processes Endure, Whereas Events Occur, dans S. Borgo, R. Ferrario, C. Masolo et L. Vieu (eds.), *Ontology Makes Sense: Essays in honor of Nicola Guarino*, Frontiers in Artificial Intelligence and Applications, 136, IOS Press, pp. 177-193, 2019.

[15] G. Kassel, Physical processes, their life and their history, *Applied Ontology*, Vol. 15, N° 2, pp. 109-133, 2020.

[16] G. Kassel, Quelle place accorder aux objets abstraits dans les ontologies fondatrices ?, dans M. Lefrançois (éd.), Actes des *32es Journées Francophones d'Ingénierie des Connaissances* (IC) *Plate-Forme Intelligence Artificielle* (PFIA), pp. 65-72, 2021.

[17] G. Kassel, Abstract events in semantics, *Philosophia*. *À paraître*.

[18] G. Kassel, In defense of epistemic ontologies. *Soumis*.

[19] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari et L. Schneider, The WonderWeb Library of Foundational Ontologies and the DOLCE ontology, WonderWeb Deliverable D18, Final Report, vr. 1.0, 2003.

[20] G.H. Merrill, Ontological realism: Methodology or misdirection?, *Applied Ontology*, Vol. 5, pp. 79-108, 2010.

[21] K. Mulligan, P. Simons et B. Smith, Truth-Makers, *Philosophy and Phenomenological Research*, Vol. 44, pp. 287-321, 1984.

[22] E. Pacherie, The content of intentions, *Mind and Language*, Vol. 15, N° 4, pp. 400-432, 2000.

[23] C. Panaccio, *Les mots, les concepts et les choses. La sémantique de Guillaume d'Occam et le nominalisme d'aujourd'hui*, Bellamin / Vrin – Analytiques, 1992.

[24] J. Plourde, Wittgenstein et les théories du jugement de Russell et de Meinong, *Dialogue*, Vol. 44, N° 2, pp. 249-283, 2005.

[25] S. Richard, Twardowski sur la vérité, *La Revue philosophique de Louvain*, Vol. 115, N° 4, pp. 619-645, 2017.

[26] G. Rosenkrantz, et J. Hoffman, The Independence Criterion of Substance, *Philosophy and Phenomenological Research*, Vol. 51, N° 4, pp. 835-853, 1991.

[27] J.R. Searle, *Intentionality*, Cambridge University Press, 1983.

[28] J. Searle, *The Construction of Social Reality*, New York: The Free Press, 1995.

[29] B. Smith, *Austrian Philosophy, Brentano's Legacy*, Chicago, Open Court, 1994.

[30] B. Smith et W. Ceusters, Ontological realism: A methodology for coordinated evolution of scientific ontologies, *Applied Ontology*, Vol. 5, pp. 139-188, 2010.

[31] A.D. Spear, W. Ceusters et B. Smith, Functions in Basic Formal Ontology, *Applied Ontology*, Vol. 11, pp. 103-128, 2016.

[32] A.L. Thomasson, Foundations for a Social Ontology, *Protosociology*, Vol. 18-19, pp. 269-290, 2003.

[33] K. Twardowski, Sur la théorie du contenu et de l'objet des représentations, dans J. English (éd.), *Husserl – Twardowski, sur les objets intentionnels (1893-1901)*, Paris, J. Vrin, pp. 85-200, 1993 ; trad., introduction et notes par J. English de *Zur Lehre vom Inhalt und Gegenstand der Vorstellungen. Eine psychologogische Untersuchung*, Vienne, Hölder, 1894.

[34] K. Twardowski, Fonctions et Formations. Quelques remarques aux confins de la psychologie, de la grammaire et de la logique, dans D. Fisette & G. Fréchette (dir.), *À l'école de Brentano. De Würzbourg à Vienne*, Paris : J. Vrin, pp. 343-283, 2007 ; trad. fr. par L. Joumier et J. Plourde de Über Gebilde und Funktionen. Einige Bermerkungen zum Grenzgebiete der Psychologie, Grammatik und Logik, dans A. Ruge (dir.), *Die Philosophie der Gegenwart*, Heidelberg: Weiss, 1911.

[35] K. Twardowski, Theory of Knowledge. A Lecture Course, dans J. Brandl et J. Voleński (eds.), *On Actions, Products and other Topics in Philosophy*, Amsterdam et Atlanta, Rodopi, pp. 182-239, 1999 ; trad. angl. par A Szylewicz, de Teoria poznania. Wyklad czterogodzinny lato 1924-1925, I. Dambska (ed)., *Archiwum historii filozofii I mysli spolecznej*, Vol. 21, pp. 241-299, 1925.

[36] N. Wilson, Facts, Events, and Their Conditions, *Philosophical Studies*, Vol. XXV, pp. 303-321, 1974.

Où sont les termes ?

Béatrice Markhoff[1], Arnaud Soulet[1]

[1] Université de Tours, LIFAT

prenom.nom@univ-tours.fr

Résumé

Pour donner une idée concise du contenu d'un graphe de connaissances, il est classique de montrer les classes et les propriétés qui y sont instanciées. Pourtant d'autres éléments peuvent informer sur ce contenu autant que les classes et les propriétés, ce sont les termes de vocabulaires contrôlés, des mots associés à des concepts. Nous présentons une étude sur le rôle et la place de ces termes dans des graphes de connaissances du Web, en particulier ceux conçus avec le CIDOC CRM. Nous expliquons les difficultés qu'il y a à les retrouver automatiquement. Nous recensons des requêtes simples pour ce faire, nous en proposons une plus complexe et nous présentons des résultats d'expérimentations sur des points d'accès SPARQL dans le domaine du patrimoine culturel, lesquels montrent que... la question reste ouverte.

Mots-clés

CIDOC CRM, graphe de connaissances, ontologie, SPARQL, thésaurus, terme, terminologie.

Abstract

To give a quick idea of a knowledge graph content, it is usual to show the classes and properties it instantiates. However, controlled vocabulary's terms can also inform about its content, as much as the classes and properties. We present a study on the role and the place of these terms in knowledge graphs, in particular those designed with the CIDOC CRM. We explain the difficulties of finding them automatically. We identify simple queries to do it, we propose a more complex one, and we present experimental results on SPARQL access points, in the cultural heritage domain. Those results show that... the question remains open.

Keywords

CIDOC CRM, knowledge graph, ontology, SPARQL, thesaurus, term.

1 Introduction

L'origine de cet article est un travail [3] sur le profilage de graphes de connaissances du Web se rapprochant de ce qui est proposé dans [16] ou [6]. Un profil consiste grossomodo en un graphe des classes et des propriétés instanciées. Par exemple pour le graphe de connaissances d'Epicherchel, contenant les objets sur lesquels ont été trouvées des inscriptions antiques à Césarée de Maurétanie, la figure 1 [1] montre dans sa partie gauche qu'il contient des instances de la classe *S19 Encounter event* et de la classe *E22 Man-Made Object* (en passant la souris sur le nom des classes on peut savoir combien) et également qu'il y a 182 triplets (sujet, prédicat, objet) où le sujet est de la classe *S19 Encounter event*, l'objet de la classe *E22 Man-Made Object* et le prédicat est la propriété *O19 has found object* [2]. Générer le profil d'un graphe de connaissances est utile pour informer sur son contenu et pour guider son exploration et son interrogation.

Pour Epicherchel comme pour les autres graphes de connaissances du portail OpenArchaeo [3], certaines propriétés ont pour objet un concept d'un vocabulaire contrôlé, dont le terme associé informe plus précisément sur le contenu du graphe que les seules classes et propriétés. Pour Epicherchel on voit en figure 1 que c'est le cas des propriétés *P2 has type*, *P101 had as general use* et *P45 consists of*. Par exemple *P101 had as general use* relie des instances de *E22 Man-Made Object* à des instances de *E55 Type* et aussi à des concepts, représentés par le noeud *autel et al*, concepts dont les termes associés sont listés lorsqu'on passe la souris sur ce noeud, comme montré dans la partie droite de la figure. Il est très utile de montrer ces termes dans un profil parce que savoir qu'il y a des milliers d'instances de *E22 Man-Made Object* dans un graphe indique juste que ce graphe contient des informations sur des objets du patrimoine culturel, mais voir en plus ces termes donne une idée plus précise, permettant de savoir dans l'exemple que ces objets relèvent de fouilles archéologiques de sites antiques du pourtour méditerranéen.

La problématique que nous montrons dans cet article est qu'il est difficile de caractériser ce genre de termes de vocabulaires contrôlés dans un graphe de connaissances et donc de les détecter automatiquement.

Pourtant ces termes sont particulièrement porteurs de connaissances : encore aujourd'hui et depuis l'origine de leur discipline, les archéologues apportent une attention particulière aux termes utilisés pour renseigner les bases de données recensant leurs découvertes. Des typologies se sont constituées, motivées par la nécessité de nommer «

1. Visible en ligne ici : `https://kgsumviz.univ-tours.fr/Home.php`
2. Ces classes et propriétés appartiennent au CIDOC CRM ou à ses extensions : `https://cidoc-crm.org`
3. `http://openarchaeo.huma-num.fr/explorateur/home`

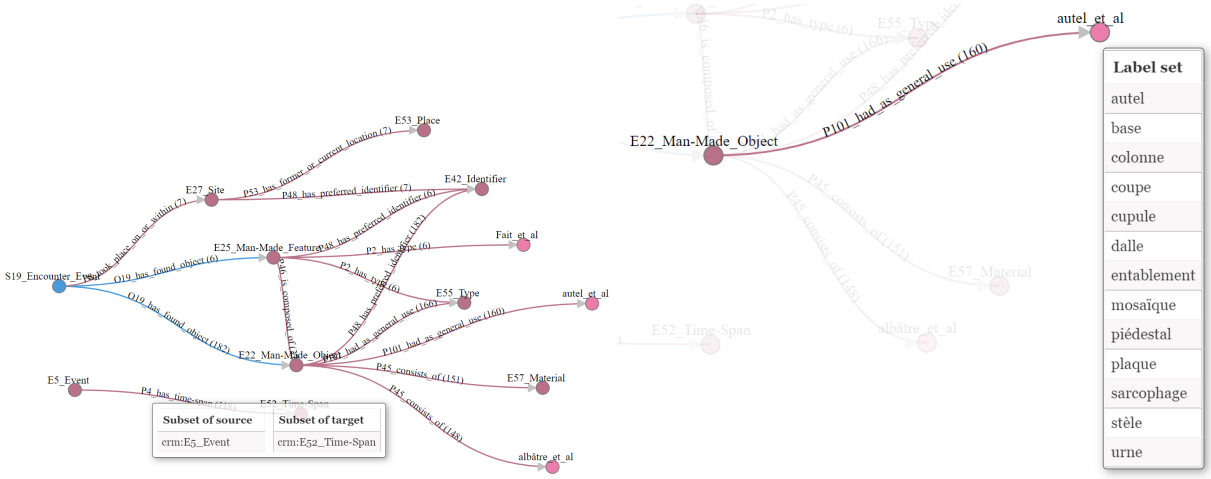


FIGURE 1 – Profil [3] du graphe Epicherchell dans OpenArchaeo.

correctement » les objets découverts, leurs fonctions, leur forme, leur constitution (matériaux), et toutes leurs caractéristiques. Il existe une tension récurrente entre la liberté de choix et de précision des termes utilisés, choix et précision qui relèvent de l'activité savante, et la nécessité de consensus, au moins à un assez haut niveau d'abstraction, pour permettre le partage d'information [10]. Le besoin d'une normalisation des termes utilisés dans les descriptions est d'autant plus ressenti avec l'ouverture des ressources via le Web, pour ce qu'il permet d'échanges et de partage. Comme l'interopérabilité sémantique passe en particulier par les termes utilisés, la communauté a mis en œuvre des systèmes en ce sens [4, 13], en parallèle de travaux sur la représentation des connaissances sous la forme d'ontologies. De façon intéressante, en archéologie comme ailleurs (par exemple dans le domaine médical avec le système d'indexation MeSH et le vocabulaire SNOMED CT), le travail sur les termes est pris en charge par des spécialistes de la gestion de l'information (documentalistes) et par des spécialistes du domaine, en parallèle ou conjointement, avec cette motivation double : d'une part *indexer pour retrouver* l'information, d'autre part *décrire* son objet d'étude et son étude elle-même. Nous revenons sur ces deux motivations distinctes dans la section 2, en donnant également un aperçu de l'importance de la production de ressources terminologiques dans le Web sémantique.

Ainsi, alors que les ontologies du Web sémantique décrivent déjà les domaines représentés via un vocabulaire structuré (Terminological Box des logiques de description), les termes sont très souvent *aussi dans les données* (Assertional Box) et c'est même une bonne pratique recommandée dans [12] : il faut utiliser des termes de vocabulaires partagés, de préférence standardisés, pour encoder les données et les métadonnées. Les auteurs de la recommandation indiquent que les bénéfices de cette bonne pratique sont : la réutilisation, la facilité de traitement, la compréhension, la confiance et l'interopérabilité. Dans ces vocabulaires les termes sont associés à des URIs, décrits de façon non ambiguë et qui peuvent avoir des labels dans différentes langues. Nous verrons en section 2 que ces vocabulaires peuvent prendre la forme d'ontologies (décrites en RDFS ou OWL) ou de thésauri (décrits en SKOS pour la plupart). Etant donné que les ontologies sont également, comme les ressources terminologiques, des supports d'interopérabilité sémantique, nous rappelons en section 3 les relations qu'elles peuvent entretenir avec les ressources terminologiques, avec un focus sur l'ontologie CIDOC CRM [2] pour le patrimoine culturel.

Ces réflexions ont pour objet de déterminer *comment détecter des termes* dans un graphe de connaissances [4]. Car autant la propriété `rdf:type` dénote une instance de classe et la 2ème position dans un triplet dénote une instance de propriété, autant détecter un terme ne repose sur aucun élément de syntaxe aussi simple. Dans la section 4, nous listons donc différentes formes de syntaxe pouvant dénoter des termes et les requêtes SPARQL correspondantes, puis nous présentons des résultats d'expérimentations sur différents points d'accès SPARQL liés au patrimoine culturel, montrant dans quelle mesure ces indices fonctionnent. Enfin, nous concluons en section 5 sur cette étude.

2 Que sont les termes dans le Web sémantique

Ce que nous appelons ressource terminologique dans le Web sémantique est conçu selon deux objectifs, l'indexation à des fins de recherche d'information et l'élaboration de terminologies.

2.1 Terminologies

Lorsqu'une communauté recherche un consensus sur les mots à utiliser pour décrire des éléments de son domaine, y compris dans des langues différentes, elle conçoit une terminologie. C'est utile en particulier pour renseigner de fa-

4. Graphe de connaissances au sens défini dans [7] qui, de façon intéressante, ne parle pas du tout de termes.

çon cohérente des champs de bases de données. En ce sens les termes sont des *données*. La norme ISO 1087 :2019 du groupe ISO/TC 37 définit une terminologie comme un ensemble de désignations utilisées dans un langage de spécialité, où une désignation représente un *concept* par un *signe* qui le dénote. L'ISO/TC 37 est également à l'origine des standards TMF (Terminological Mark-up Framework) et LMF (Lexical Mark-up Framework) qui ont inspiré l'ontologie OntoLex-lemon [5] (représentation des propriétés morpho-syntaxiques des entrées lexicales et de leur sens) et son extension pour la terminologie en cours de définition, Termlex [6], dédiée à la documentation des informations sur les termes. Cette proposition permet de représenter clairement l'interface entre syntaxe (signe) et sémantique à l'aide du concept LexicalSense, qui peut faire le lien vers une ontologie dans laquelle le concept est décrit [7]. La question de ce lien entre forme lexicale et concept pour la description de termes fait également l'objet de proposition de système onto-terminologique [14].

Cependant en général dans le Web, certaines terminologies sont réalisées juste sous forme d'ontologie, comme le Dublin Core [8], et d'autres sont réalisées juste avec SKOS ou SKOS-XL [9], comme le AAT du Getty [10]. Pourtant SKOS a une expressivité limitée pour cet usage, puisque c'est une ontologie pour définir des thésauri, taxonomies, schémas de classification ou systèmes de vedettes-matières, utilisés dans des systèmes documentaires à des fins d'indexation et de recherche d'information.

2.2 Indexation et recherche d'information : les thésauri

Un thésaurus est un vocabulaire contrôlé et structuré dans lequel les concepts sont représentés par des termes, où des relations entre les concepts sont explicitées et où les termes préférés sont accompagnés d'entrées de synonymes ou de quasi-synonymes (voir ISO 25964-1 sections 2.62 thésaurus et 2.35 thésaurus multilingue). Dans ce cadre un concept est une unité de pensée et un terme est un mot ou une expression utilisée pour étiqueter un concept (voir ISO 25964-1 sections 2.11 Concept et 2.61 Terme). Ces définitions se rapprochent de celles d'une terminologie et justifient le terme de « ressource terminologique » pour parler aussi bien de terminologie que de thésaurus. Une différence entre thésauri et terminologies réside toutefois dans leur vocation (nous avons vu que celle des terminologies est le consensus sur les désignations utilisées dans un langage de spécialité, soit l'interopérabilité sémantique). L'objectif principal des thésauri est d'indexer et de retrouver des éléments (souvent des documents) en fonction de leur contenu : « le document D traite le sujet C ». Le thésaurus sert alors de structure d'accès, sachant que les déclarations de synonymie entre termes d'une part, et d'autre part la relation hiérarchique entre concepts, permettent au système de recherche d'information d'élargir ou de restreindre les requêtes. Dans ce but, la hiérarchie utilisée dans un thésaurus couvre la relation de subsomption, la relation de partition, parfois aussi la relation d'instanciation, fusionnées en une relation hiérarchique unique dans certains thésaurus pour répondre à cette définition fondée sur l'usage : « le concept A est plus large que le concept B » si « dans toute recherche de A, les articles traitant de B devraient être retournés ». Les grands thesaurus sont organisés en facettes, qui regroupent des hiérarchies de concepts pour faciliter la recherche d'information. Comme le note [9], la structure des thésauri n'est pas utile à des raisonnements plus généraux, contrairement aux ontologies.

2.3 Ressources terminologiques dans le Web

On trouve des ressources terminologiques sous la forme de thésauri ou d'ontologies dans le Web sémantique, les deux étant supports d'interopérabilité sémantique, les premiers au niveau données et les secondes au niveau métadonnées. Les ressources terminologiques sont plus massivement réutilisées que des ontologies conçues pour représenter un domaine de connaissance, dans la mesure où le consensus nécessaire à leur réutilisation porte sur les termes (et leur définition en contexte), et non pas sur la question plus complexe de la manière dont la représentation du domaine est organisée et structurée (ontologie). Il est plus simple de choisir un terme pertinent dans un thésaurus que de comprendre et réutiliser une ontologie, sauf pour les plus simples d'entre elles comme FOAF et le Dublin Core, qui sont des ressources terminologiques. Ainsi, il existe de nombreuses et, pour certaines, très grandes ressources terminologiques dans le Web. Par exemple dans le domaine biomédical UMLS [11] rassemble des concepts de plusieurs dizaines de terminologies, MeSH [12] (défini par la bibliothèque nationale de médecine des Etats-Unis) permet d'indexer les répertoires d'articles Medline et PubMed, quand SNOMED CT [13] rassemble plusieurs centaines de milliers de concepts utilisés dans des environnements cliniques, en particulier pour les dossiers patients. De même dans le domaine environnemental, AGROVOC [14] regroupe plus de 38 000 concepts de l'alimentation, agriculture, pêche, foresterie, etc. auxquels sont associés plus de 800 000 termes dans 40 langages. Dans le domaine du patrimoine culturel, le Backbone thesaurus de DARIAH [15] est une initiative pour l'agrégation et la maintenance de vocabulaires construits dans des communautés, comme les PACTOLS déjà cités pour l'archéologie, mais c'est surtout le vocabulaire AAT

5. https://www.w3.org/2016/05/ontolex/
6. https://www.w3.org/community/ontolex/wiki/Terminology
7. Voir http://anr-sesames.map.cnrs.fr/onto/shart/index.html pour un exemple d'utilisation
8. Voir https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
9. https://www.w3.org/TR/skos-reference/
10. Art and Architecture Thesaurus : https://www.getty.edu/research/tools/vocabularies/aat/
11. Unified Medical Language System : https://www.nlm.nih.gov/research/umls/index.html
12. Medical Subject Headings : https://www.ncbi.nlm.nih.gov/mesh
13. Systematized Nomenclature Of Medicine Clinical Terms : https://www.snomed.org/
14. https://www.fao.org/agrovoc/
15. https://www.backbonethesaurus.eu/

du Getty, déjà cité également, qui est utilisé et avec lequel les thésauri locaux sont alignés, par exemple dans EUROPEANA [16] ou dans la plateforme ARIADNE+ [17]. Celle-ci organise les recherches possibles selon trois axes Quand-Où-Quoi : le Getty AAT est utilisé dans l'axe Quoi, pour décrire ce qui est recherché [18]. Il est intéressant de noter ici que pour des axes de recherche comme Quand (périodes historiques), Où (lieux) et Qui (personnes, organisations), aussi bien dans la plateforme ARIADNE+ que dans OpenArchaeo, des URIs issus de listes d'autorité sont également utilisés, mais il s'agit alors *d'entités nommées*, qui se réfèrent chacune à un élément unique, et non plus des *termes* qui, eux, *sont des « universaux » [11] au même titre que les classes et les propriétés d'une ontologie*, même s'ils sont des instances de `skos:Concept`. Cette dernière remarque sera exploitée dans la section 4.

3 Relations avec les ontologies

3.1 Ontologie versus thésaurus

Il n'est pas question ici de définir précisément ce qu'est une ontologie dans le Web comme le font les auteurs de [9], rappelons simplement que c'est un modèle formel, et consensuel, d'une conceptualisation d'un domaine de connaissances. Il est constitué de description intensionnelle (TBox) et extensionnelle (ABox, rassemblant les instances). Il comporte un ensemble d'entités, de relations et d'axiomes pour les décrire. Une entité y est vue comme une classe qui a des instances, le lien de subsomption étant que si A subsume B alors toute instance de B est aussi instance de A. Les relations sont décrites par leur domaine et codomaine et il y a également un lien de subsomption entre relations. Selon les besoins d'autres précisions peuvent être spécifiées. L'ensemble de ces déclarations est exploitable automatiquement par des raisonneurs. Ainsi, une bonne pratique recommandée dans [5] pour modéliser des ressources et leurs contenus est d'affecter à la ressource une propriété `hasTopic` dont l'objet est instance d'une classe représentant ce contenu. Sachant que les classes représentant les contenus sont organisées selon leurs liens de subsomption, cela permet de bénéficier directement du raisonneur : demander des ressources portant sur un métal revient à demander aussi celles portant sur toutes les sortes de métal subsumées par la classe Metal. On reconnaît là la vocation d'un thésaurus, mais dans un thésaurus c'est un système ad hoc qui réalise cette généralisation de la requête, pas un raisonneur générique. La question de transformer les connaissances contenues dans un thésaurus en ontologie pour les exploiter automatiquement par un raisonneur est abordée de longue date [1, 8]. Elle n'est pas simple, notamment du fait que la relation broader-narrower d'un thésaurus peut être aussi bien une relation de subsomption qu'une relation de partition, et que la relation associative générique souvent présente dans un thésaurus est également complexe à transposer automatiquement dans une ontologie. De plus, il faut décider quels concepts du thésaurus sont représentés par des classes et lesquels sont des données (instances). Pour autant il existe des versions de SNOMED CT et d'AGROVOC sous forme d'ontologies OWL. Dans le domaine du patrimoine culturel, la communauté qui définit et maintient l'ontologie CIDOC CRM a une politique claire pour l'usage de descriptions terminologiques conjointement avec l'ontologie, que nous présentons succinctement dans la section suivante.

3.2 CIDOC CRM

CIDOC CRM est une ontologie conçue pour supporter l'interopérabilité sémantique de ressources numériques du patrimoine culturel, administrée par un groupement d'intérêts (SIG) depuis les années 80. La question des termes est abordée dès l'introduction du document qui la définit, avec ce titre : « About Types ». Une classe particulière, `E55 Type`, est destinée à regrouper les termes de thésaurus et vocabulaires contrôlés utilisés pour caractériser et classifier les instances de classes de CIDOC CRM. La classe la plus haute dans la hiérarchie de subsomption, `E1 CRM Entity` est le domaine de la propriété `P2 has type`, dont le codomaine est `E55 Type`. Ainsi, chaque classe de CIDOC CRM (à l'exception de `E59 Primitive Value`), hérite de la propriété `P2 has type`, ce qui fournit un mécanisme général pour spécialiser la classification des instances dans un graphe de connaissances utilisant CIDOC CRM à n'importe quel niveau de détail, en établissant un lien avec des sources externes, thésaurus ou ontologies. Pour classifier ainsi, il est possible d'implémenter le concept soit comme une *sous-classe* étendant le système de classes de CIDOC CRM, soit comme une *instance* de `E55 Type`. Selon les principes de construction de CIDOC CRM, une nouvelle sous-classe ne doit être créée que si le concept est suffisamment stable et associé à des propriétés supplémentaires explicitement modélisées qui lui sont propres. Sinon, une instance de `E55 Type` doit être choisie. Ce traitement cohérent des connaissances de nature terminologique renforce la capacité de CIDOC CRM à *servir de pivot d'intégration* de connaissances relevant du patrimoine culturel. En plus d'être une interface vers des thésaurus et des systèmes de classification externes, `E55 Type` est une sous-classe de `E28 Conceptual Object`. `E55 Type` et ses sous-classes héritent donc également de toutes les propriétés de cette super-classe. L'une des réflexions en cours au sein du SIG CRM porte encore sur les terminologies dans le domaine du patrimoine culturel. Une note récente de Martin Doerr précise ceci : les classes sans propriété sont modélisées comme des instances de `E55 Type`, c'est-à-dire comme des *données*. Ces données définissent un vocabulaire, sachant que les vocabulaires non standardisés sont un outil important de la recherche dans toutes les sciences et les humanités. A des fins d'interopérabilité le SIG CRM recommandera cependant dans un document distinct de la définition de CIDOC CRM certains termes, mais uniquement ceux considérés comme importants pour certaines distinctions ontologiques précises, et suffisamment univoques pour être fixés

16. `https://pro.europeana.eu/page/europeana-aat`
17. `https://ariadne-infrastructure.eu/Portal/`
18. Periodo (`perio.do`) est utilisé pour l'axe Quand et Geonames (`geonames.org`) pour l'axe Où.

comme standards. Ces termes pourront être liés ou intégrés en tant que termes plus larges ou plus étroits dans les vocabulaires utilisés par l'utilisateur, d'une manière compatible avec la signification des classes de CIDOC CRM. En outre, le SIG CRM recommandera l'utilisation de certains vocabulaires standardisés dans les cas où il existe une pratique internationale, comme pour les unités de mesure, les codes de pays, etc. Lors de la création des graphes de connaissances de OpenArchaeo, évoqués en introduction, le principe de l'utilisation de la classe `E55 Type` a été appliqué pour articuler l'utilisation de l'ontologie CIDOC CRM avec celle du thésaurus PACTOLS. Pour attribuer des labels aux entités, la question s'est posée d'utiliser `rdfs:label` ou `skos:label` et le choix suivant a été fait par les archéologues : s'il s'agit d'une entité intermédiaire (présente dans le graphe de connaissances juste pour respecter CIDOC CRM) alors un `rdfs:label` peut lui être associé si besoin, mais s'il s'agit d'une entité dont le label est un terme alors un `skos:prefLabel` lui est attaché, même si cette entité n'est pas (encore) un concept d'un thésaurus [19]. Le but étant de *marquer une intention* d'utiliser ce label comme un terme (en attendant qu'il soit dans PACTOLS).

4 Détection de termes dans un graphe de connaissances

4.1 Indices

Nous avons tenté de suivre quatre indices pour la détection des termes, résumés dans la table 1, avec trois principaux critères de comparaison. D'abord, l'objectif est d'avoir une approche avec peu de faux positifs (précision élevée) et couvrant un maximum de termes (rappel élevé). Ensuite, une approche est d'autant plus intéressante qu'elle requiert peu de connaissances d'autres ressources a priori (prérequis faible).

Classe répertoriée Parfois, les concepts dénotant des termes sont explicitement déclarés dans le graphe analysé comme instances d'une classe qui est connue pour représenter des ressources terminologiques, comme `skos:Concept`, `skos:Collection`, `crm:E55 Type` pour CIDOC CRM ou `ontolex:LexicalConcept` pour OntoLex-lemon et Termlex. En effet pour les graphes de connaissances utilisant CIDOC CRM et suivant ses recommandations (cf. section 3.2), les instances de `crm:E55 Type` (et de ses éventuelles sous-classes) doivent être des concepts dénotant des termes. Cet indice est sûr. Rien que la classe `skos:Concept` permet(trait) de couvrir de nombreux termes de par son usage fréquent. Hélas, le graphe de connaissances interrogé ne contient pas souvent cette déclaration, présente uniquement dans le thésaurus (c'est notamment le cas pour OpenArchaeo), même si cela peut arriver (Cultura Italia contient les définitions des classes, propriétés et concepts SKOS). Cet indice demande de connaître l'ontologie de référence du domaine (par exemple CIDOC CRM pour savoir que la classe `E55 Type` a pour instances des concepts de terminologies).

Propriété répertoriée Les sujets de certaines propriétés ont une forte chance d'être des concepts dénotant des termes. Nous avons d'abord pensé que c'est le cas de `skos:prefLabel`, on pourrait même espérer que ce soit un indice universel de la même manière que `rdfs:type` identifie une instance de classe. Il n'en est rien en pratique car, même si cette propriété est effectivement largement adoptée, aucune convention n'encadre son usage : sa définition ne contraint pas son domaine et il n'existe pas de bonnes pratiques clairement énoncées à notre connaissance. De ce fait la précision de cet indice avec cette propriété est faible. Le rappel est potentiellement élevé toutefois il peut arriver que, plutôt qu'un `skos:prefLabel`, ce soit un `skosxl:prefLabel` qui soit utilisé, lorsque c'est une entité lexicale qui est associée au concept, plutôt qu'une chaîne de caractères. D'autres propriétés relevant d'une description de terminologie peuvent être testées, comme d'autres propriétés de SKOS, celles de SKOS-XL, ou `ontolex:isEvokedBy` pour OntoLex-lemon et Termlex, ou les propriétés `P127 has broader term` et `P150 defines typical parts of` pour le CIDOC CRM. Mais, comme pour les classes répertoriées, ces propriétés sont rarement présentes dans le graphe analysé, mais plutôt dans le graphe où est définie la terminologie.

Préfixe répertorié Une méthode relativement naïve est de détecter les URIs correspondant à des thésauri connus. Par exemple, toutes les URIs débutant par `https://ark.frantiq.fr/` appartiennent au thésaurus PACTOLS [20]. La précision de cette approche est très élevée, mais sa couverture dépend du fait que le graphe de connaissances ciblé utilise des thésauri connus ou pas. Cette approche ne permet pas la découverte de nouveaux thésauri et elle est donc peu adaptée aux graphes de connaissances nouveaux.

Universaux filtrés En métaphysique, les universaux sont « des termes généraux qui semblent désigner ce qui est commun entre diverses choses » [11], ces termes étant utilisés pour désigner et comprendre les caractéristiques communes des *entités particulières* [15]. En ce sens, les concepts dénotant des termes sont donc des *entités universelles* au même titre que les classes ou les propriétés [21]. Nous avons cherché à détecter l'ensemble des entités universelles d'un graphe de connaissance, puis d'en retirer les classes et les propriétés. Notre méthode de détection repose sur l'idée clé qu'une *entité universelle* n'est jamais sujet d'une assertion dans laquelle une *entité particulière* est objet : une entité universelle définit une entité particulière puisqu'elle représente une caractéristique commune à

19. La propriété `skos:prefLabel` n'a pas de domaine défini et peut s'appliquer à tout type d'objet : `https://www.w3.org/TR/2009/REC-skos-reference-20090818/#L1541`

20. Il faut évidemment que tous les concepts du thésaurus partagent le même préfixe et que ce préfixe désigne uniquement des concepts dénotant des termes. Cela pose problème pour SNOMED CT.

21. Les concepteurs d'ontologies et de terminologies du Web sémantique se situeraient côté anti-réalistes, pour qui il n'existe dans la réalité que des choses particulières (ABox), les universaux se trouvant soit dans le langage (nominalistes : terminologies), soit dans l'esprit (conceptualistes ou idéalistes : TBox).

Indice	Description	Précision	Rappel	Prérequis
Classe répertoriée	Instance d'une classe dédiée à la terminologie comme `skos:Concept`, `crm:E55_Type`, etc.	très élevée	variable	élevé
Propriété répertoriée	Sujet d'une propriété en principe dédiée à la terminologie comme `skos:prefLabel`, etc.	faible	variable	élevé
Préfixe répertorié	Préfixe d'URI correspondant à un thésaurus répertorié, comme PACTOLS ou AAT du Getty	très élevée	élevé	très élevé
Universaux filtrés	Entité décrite uniquement par des littéraux ou des universaux, qui n'est ni une classe, ni une propriété	variable	élevé	faible

TABLE 1 – Quatre indices pour la détection des termes.

des entités particulières, mais elle n'est pas elle-même définie par une entité particulière. Nous recherchons donc des entités qui sont des sujets de triplets, mais pas de triplet dont l'objet est une instance de classe (entité particulière). Cela se traduit par la requête SPARQL suivante :

```
1  SELECT DISTINCT ?s
2  WHERE {
3    ?s ?p ?o .
4    FILTER NOT EXISTS {
5      ?s ?other_p ?other_o .
6      ?other_o a ?co .
7      FILTER(STR(?co)!="http://.../skos/core#Concept"
8      && !STRSTARTS(STR(?other_p),"http://.../skos"))
9    } .
10   FILTER (!ISBLANK(?s)) .
11   FILTER NOT EXISTS {?another_s ?s ?another_o} .
12   FILTER NOT EXISTS {?instance a ?s}
13 }
```

Cette requête recherche les entités universelles en prenant toutes les entités qui sont le sujet `?s` d'au moins une assertion (ligne 3), qui ne réfèrent à aucune entité particulière (ici, instance d'une classe avec la ligne 5 et 6) et qui ne sont pas des *blank nodes* (ligne 10). De plus, on ne souhaite pas éliminer les cas où la propriété `?other_p` appartient à l'ontologie `SKOS` (ligne 8) afin de respecter l'indice « Propriété répertoriée », ni ceux où la classe `?co` est `skos:Concept` (ligne 7), pour respecter l'indice « Classe répertoriée ». Enfin, les lignes 11 et 12 filtrent les termes en empêchant respectivement que l'entité `?s` soit une propriété et une classe. La force de cette proposition est de nécessiter peu de connaissances prérequises sur les graphes à analyser [22], ce qui est particulièrement adapté pour la détection de termes dans des graphes de connaissances nouveaux. Son rappel risque d'être plus élevé que les autres approches mais elle risque de détecter des faux positifs, à savoir : des entités particulières reliées à aucune autre entité particulière d'une part, et d'autre part des entités universelles qui ne seraient ni des concepts dénotant des termes, ni des classes/propriétés.

4.2 Expérimentations

Nous reportons dans cette section nos premières expérimentations correspondant à l'application des quatre méthodes de détection sur dix graphes de connaissances disponibles via des points d'accès publics : ADS [23] qui regroupe les connaissances en archéologie du Royaume-Uni, Cultura Italia [24] qui contient des connaissances sur des collections de musées et galeries italiennes, et 8 graphes de OpenArchaeo [25], portail sémantique d'une communauté d'archéologues regroupée au sein du consortium MASA [26] de la TGIR Huma-Num. Le tableau 2 donne quelques informations statistiques (nombre d'entités et d'assertions) et il présente les résultats obtenus, avec le nombre de concepts dénotant des termes découverts pour chaque méthode. Ces résultats sont obtenus en testant les classes, propriétés et thésauri cités dans la colonne Description du tableau 1. A noter qu'un usage immodéré de la propriété `skos:prefLabel` rend illusoire l'utilisation de la méthode « Propriété répertoriée » pour les graphes de OpenArchaeo. De la même façon, il y a dans Cultura Italia 866 instances distinctes de `skos:Concept` et 358 492 instances distinctes de `E55 Type` : leur analyse montre que 1 600 d'entre eux sont bien des termes d'un thesaurus, lequel est contenu dans le graphe de connaissances, mais que tout le reste correspond à des URIs décrivant des *valeurs littérales de dimensions*, comme diamètre, hauteur, etc. et de combinaisons de telles dimensions, comme par exemple `ci:format/cm-diametro-29-peso-20-6` [27]. Là encore nous pouvons conclure à un usage immodéré, voire injustifié de la classe `E55 Type`.

Les concepts dénotant des termes trouvés avec la méthode des universaux filtrés sont trouvés avec les méthodes « Classe répertoriée » ou « Propriété répertoriée » en ce qui concerne les graphes ADS et Cultura Italia. Pour les graphes de OpenArchaeo, les concepts dénotant des termes retrouvés appartiennent aux thésaurus PACTOLS et Getty. Pour ADS, les concepts dénotant des termes retrouvés appartiennent à différents thésauri (archaide, romanamphorae, etc.) qui nous étaient inconnus. Dans plusieurs graphes de OpenArchaeo, cette méthode ne donne rien parce que les concepteurs ont associé aux concepts du thésaurus PACTOLS un autre terme que dans le thesaurus en utilisant la propriété `skosxl:altLabel`. Dans un objectif d'interopérabilité, ils utilisent systématiquement un concept du thésaurus PACTOLS pour certaines propriétés du graphe, mais lorsque dans le jeu de données d'origine, fourni par

22. On peut en ajouter davantage au niveau des lignes 7 et 8.
23. `http://data.archaeologydataservice.ac.uk/query/`
24. `http://dati.culturaitalia.it/sparql`
25. `http://openarchaeo.huma-num.fr/federation/sparql`
26. `https://masa.hypotheses.org/`
27. ci <`http://dati.culturaitalia.it/resource/`>

Graphe	Entités	Assertions	Classe répertoriée	Propriété répertoriée	Préfixe répertorié	Universaux filtrés
ADS	214 155	1 547 192	1863	1588	-	1251
Cultura Italia	9 951 821	41 901 551	359358	0	981	749
arsol	105 054	669 099	75	-	150	0
chronique	98 837	557 724	114	-	118	1
epicherchell	706	3 945	18	-	32	0
iceramm	7 456	44 538	42	-	581	0
kition-pervolia	6 490	32 714	93	-	114	94
outagr	30 313	115 462	400	-	584	414
rita	12 366	76 885	518	-	479	0
solidar	9 848	48 460	94	-	159	127

TABLE 2 – Détection des concepts termes, en utilisant les différentes méthodes.

	Faux positifs				
Graphe	Part.	Prop.	Classes	Types	Autre
ADS	539	0	0		0
Cultura Italia	0	2	4	59	0
aerba	0	0	0	0	0
arsol	0	0	0	0	0
chronique	0	0	0	0	1
epicherchell	0	0	0	0	1
iceramm	0	0	0	0	1
kition-pervolia	2140	0	0	0	1
outagr	14121	0	0	0	1
rita	0	0	0	0	1
solidar	1550	0	0	0	0

TABLE 3 – Analyse des FP avec le filtrage des universaux

les chercheurs, ce sont d'autres mots qui sont utilisés ils les ajoutent avec `skosxl:altLabel`. Ils ont choisi d'utiliser la propriété SKOS-XL pour dénoter que ces mots ne viennent pas de PACTOLS mais de leur jeu de données original. Le problème pour notre méthode des universaux filtrés est que cette propriété prend pour objet une instance de la classe `skosxl:Label` et que les créateurs déclarent dans le graphe que ces objets sont de cette classe. De plus, aussi bien ces objets que les concepts de PACTOLS ou AAT Getty sont systématiquement déclarés instances de `owl:NamedIndividual` par un effet de bord d'une utilisation du logiciel Protégé, que les créateurs pensent sans conséquence pour leur exploitabilité. Pourtant, déclarer que ces entités universelles que sont les concepts de thésaurus sont des « named individual » modifie significativement le regard d'un point de vue philosophique. La méthode des universaux filtrés est ainsi tributaire d'aléas dans le processus de création des graphes.

Nous analysons maintenant les résultats retournés par cette méthode qui ne sont pas des concepts dénotant des termes. Le tableau 3 détaille ces faux positifs pour la méthode fondée sur le filtrage des universaux. La précision de cette approche, vérifiée en examinant tous les résultats obtenus, est très variable : 69,9% pour ADS, 92,0% pour Cultura Italia et seulement 3,4% sur les graphes de OpenArchaeo.

Pour ADS, les faux positifs correspondent à des entités particulières, URL de documents. Pour Cultura Italia, certaines propriétés non-utilisées et classes non-instanciées ont été retournées étant donné que ce graphe comporte toutes les définitions des classes et propriétés des ontologies utilisées. Les autres faux positifs correspondent à des typages de littéraux (entiers, chaîne de caractères,...), qui peuvent aussi être considérés comme des entités universelles et pourraient être retirées en complétant notre requête SPARQL. Pour les graphes de OpenArchaeo, dans les faux positifs, seulement 3 réponses correspondent à des entités universelles (noms de graphes) et le reste correspond à des entités particulières réparties en deux catégories. La première représente des entités X, qui existent dans le jeu de données d'origine mais ne sont présentes dans aucune page web où ce jeu de données est présenté (autrement il s'agit de l'URL de la page web qui est utilisé, terminé par #X car une même page web peut montrer plusieurs entités X). La seconde catégorie contient des URLs vers d'autres graphes, par exemple vers des entités de GeoNames. Pour ces entités-là, notre requête aurait fonctionné si elle avait été appliquée conjointement sur les graphes de OpenArchaeo et sur ceux dont sont issues les entités sélectionnées.

Le tableau 2 montre donc qu'aucune des méthodes imaginées ne fonctionne parfaitement. Les propriétés ou classes répertoriées peuvent servir mais leur utilisation est délicate si les graphes de connaissances les utilisent pour autre chose que des éléments de terminologies ou de thésauri. La méthode des universaux filtrés est également dépendante de choix de conception des graphes, qui restent très libres en l'absence de conventions bâties sur un consensus. Il n'est pas toujours possible d'identifier les concepts dénotant des termes via le préfixe de leur URI, mais lorsque cela est possible cette option s'avère la meilleure. Les autres méthodes peuvent permettre d'identifier des préfixes de thésauri ou de terminologies utilisés dans les graphes.

5 Conclusion

Dans cet article, nous nous sommes intéressés à l'usage de terminologies dans les graphes de connaissances du Web, en focalisant sur le domaine du patrimoine culturel où elles viennent compléter l'ontologie CIDOC CRM. Après avoir rappelé les raisons d'exister des terminologies, nous avons noté qu'elles se manifestent dans le Web sémantique soit

comme des ontologies, soit comme des thésauri, définis soit avec SKOS, soit avec d'autres vocabulaires. Dans le premier cas (ontologies) elles apparaissent dans les profils de graphes de connaissances qui montrent les classes et les propriétés utilisées, ce qui répond à notre motivation initiale, qui est de montrer dans le profil d'un graphe de connaissances des éléments de terminologies qu'il contient. Dans le deuxième cas (thésauri), il est compliqué de les détecter car pour l'heure dans le Web sémantique la définition et l'usage des thésauri ne sont pas cadrés par des définitions ou par des usages standards précisés dans des guides de bonnes pratiques, à l'image de la définition et l'usage d'ontologies. Nous avons testé plusieurs méthodes pour identifier des éléments de ressources terminologiques dans des graphes de connaissances, avec des résultats préliminaires sur trois points d'accès SPARQL offrant des graphes liés au patrimoine culturel. Dans le cadre du patrimoine culturel, pour des graphes conçus avec le CIDOC CRM en suivant ses principes pour l'utilisation conjointe de terminologies, il est possible de trouver des termes avec l'une ou l'autre des trois premières méthodes, mais ce n'est pas généralisable. Nous avons proposé et testé une méthode plus agnostique, qui échoue toutefois dans son état actuel à distinguer les particuliers des universels dans la plupart des graphes testés. Une partie des raisons de ces échecs révèle des anomalies de conception et pourrait être résolue en modifiant les graphes, une autre pourrait l'être en raffinant la méthode. Ce sont-là les deux perspectives que nous allons explorer.

Remerciements

Ce travail est financé par l'ANR-18-CE38-0009 SESAMES. Les auteurs remercient les relecteurs anonymes et Thomas Francart et Yannick Duthé pour leurs critiques et suggestions.

Références

[1] Fabien Amarger, Catherine Roussey, Jean-Pierre Chanet, Olivier Haemmerlé, and Nathalie Hernandez. Etat de l'art : Extraction d'information à partir de thésaurus pour générer une ontologie. In *INFORSID*, pages 29–44, 2013.

[2] Chryssoula Bekiari, George Bruseker, Martin Doerr, Christian-Emil Ore, Stephen Stead, and Athanasios Velios. Definition of the CIDOC Conceptual Reference Model. last official version : 7.1.1. *ICOM/CIDOC Documentation Standards Group. CIDOC CRM SIG*, 2021.

[3] Lamine Diop, Arnaud Giacometti, Béatrice Markhoff, and Arnaud Soulet. TTProfiler : Computing Types and Terms Profiles of Assertional Knowledge Graphs. In *Proceedings of the Semantic Web and Ontology Design for Cultural Heritage workshop co-located with (BOSK 2021)*, volume 2949 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021.

[4] FRANTIQ. Le thésaurus pactols. `https://www.frantiq.fr/pactols/le-thesaurus/`, 2020. Accessed on 2022-28-02.

[5] Fabien Gandon, Catherine Faron-Zucker, and Olivier Corby. *Le Web sémantique*. Dunod, Paris, France, 2012.

[6] François Goasdoué, Pawel Guzewicz, and Ioana Manolescu. RDF graph summarization for first-sight structure discovery. *VLDB J.*, 29(5) :1191–1218, 2020.

[7] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutiérrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan F. Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge graphs. *ACM Comput. Surv.*, 54(4) :71 :1–71 :37, 2021.

[8] D. Kless, L. Jansen, and S. Milton. A content-focused method for re-engineering thesauri into semantically adequate ontologies using owl. *Semantic Web*, 7(5) :543–576, 2016.

[9] D. Kless, S. Milton, E. Kazmierczak, and J. Lindenthal. Thesaurus and ontology structure : Formal and pragmatic differences and similarities. *Journal of the Association for information science and technology*, 66(7) :1348–1366, 2015.

[10] Marion Lamé, Perrine Pittet, Federico Ponchio, Béatrice Markhoff, and Emilio M. Sanfilippo. Heterotoki : non-structured and heterogeneous terminology alignment for digital humanities data producers. In *Workshop on Open Data and Ontologies for Cultural Heritage co-located with CAiSE, ODOCH@CAiSE 2019*, volume 2375 of *CEUR Workshop Proceedings*, pages 37–48. CEUR-WS.org, 2019.

[11] Bruno Langlet. Universaux (GP), dans Maxime Kristanek (Dir.), l'Encyclopédie philosophique. `https://encyclo-philo.fr/universaux-gp`, 2019. Consulté le 15/03/2022.

[12] Bernadette Farias Lóscio, Caroline Burle, and Newton Calegari. Data on the web best practices (w3c recommendation 31 january 2017) : Best practice 15. `https://www.w3.org/TR/dwbp/`, 2017. Accessed on 2022-28-02.

[13] Emmanuelle Perrin. Thésaurus et interopérabilité des données archéologiques : le projet hyperthesau. *Humanités numériques [en ligne]*, 4, 2021.

[14] Christophe Roche and Maria Papadopoulou. Terminology and ontology for digital humanities : The case of ancient greek dress. *Humanités numériques*, 2, 2020.

[15] Bertrand Russell. *On the relations of universals and particulars*. Lulu Press, Inc, 2015.

[16] Blerina Spahiu, Riccardo Porrini, Matteo Palmonari, Anisa Rula, and Andrea Maurino. ABSTAT : Ontology-Driven Linked Data Summaries with Pattern Minimalization. In *The Semantic Web - ESWC 2016 Satellite Events, Revised Selected Papers*, pages 381–395. Springer, 2016.

Négociation de contenu sur le Web: un état de l'art

Y. Taghzouti[1], A. Zimmermann[1], M. Lefrançois[1]

[1] Mines Saint-Étienne, Univ Clermont Auvergne, INP Clermont Auvergne, CNRS, UMR 6158 LIMOS, F - 42023 Saint-Étienne France

yousouf.taghzouti@emse.fr, antoine.zimmermann@emse.fr, maxime.lefrancois@emse.fr

Résumé

L'ouverture et l'accessibilité du Web a largement contribué à son adoption à l'échelle mondiale. L'identifiant uniforme de ressource (URI) est la pierre angulaire de l'identification des ressources sur le Web. Une ressource sur le Web peut être décrite de nombreuses façons, ce qui peut rendre difficile pour un utilisateur de trouver une représentation adéquate. Cela a motivé une recherche fructueuse sur la négociation de contenu (NC) afin de satisfaire les besoins de l'utilisateur de manière efficace. Notre objectif de recherche est de présenter un état de l'art qui inclut (1) une identification des caractéristiques des scénarios de la NC (styles, dimensions et moyens de transmission des contraintes), (2) une comparaison et une classification de quelques contributions existantes, (3) des cas d'utilisation pas encore couverts dans l'état de l'art, (4) une suggestion de directions de recherche pour les travaux futurs. Les résultats de l'état de l'art montrent que le problème de la NC est pertinent et loin d'être résolu.

Mots-clés

Négociation de contenu, Web sémantique, contrainte.

Abstract

The openness and accessibility of the Web has contributed greatly to its worldwide adoption. Uniform Resource Identifier (URI) is the cornerstone of resource identification on the Web. A resource on the Web can be described in many ways. The large number of ways to describe a resource can make it difficult for a user to find an adequate representation. This situation has motivated fruitful research on content negotiation (CN) to satisfy user requirements efficiently and effectively. We focus on the topic of CN, and our goal is to present a comprehensive state of the art that includes (1) identifying the characteristics of CN scenarios (styles, dimensions, and means of conveying constraints), (2) comparing and classifying some existing contributions, (3) identifying use cases that the current state of CN struggles to address, (4) suggesting research directions for future work. The results of the state of the art show that the problem of CN is relevant and far from being solved.

Keywords

Content negotiation, semantic web, constraint.

1 Introduction

Ouvert, distribué, accessible et hétérogène sont quelques-unes des caractéristiques fondamentales du Web [5]. Le fait que chacun puisse accéder au Web de n'importe où dans le monde a grandement contribué à son développement et à son enrichissement grâce à son ouverture. Cependant, cela a eu l'effet indésirable d'avoir une abondance de ressources Web, et par conséquent des difficultés à fournir le meilleur contenu pour chaque utilisateur ; un exemple simple est celui de deux personnes parlant des langues différentes accédant à la même ressource. Le serveur avec la ressource devrait être capable de fournir à chaque utilisateur une version compréhensible. Pour remédier à cela, une solution a été imaginée dès le départ, avec une couche de négociation entre le client et le serveur [3]. Elle est abordée plus tard dans le document Architecture of the World Wide Web comme l'un des composants essentiels de la conception Web [13, section 3.2].

La négociation, en tant que concept, est une communication aller-retour destinée à atteindre un accord lorsque deux ou plusieurs parties ont des intérêts communs et d'autres opposés [10, p. 1]. Appliquée au Web, elle devient alors la "négociation de contenu" (NC) : le mécanisme permettant de sélectionner la représentation appropriée lors du traitement d'une requête. Dans HTTP, on peut exprimer et transmettre des contraintes appelées préférences [8, section 5.3]. Et avec cela, en plus de trouver et de transmettre des informations, il est possible de sélectionner des formats et des langages plus spécifiques.

Avec l'ère du mobile, un nouveau défi est apparu, puisque le contenu déjà disponible était conçu pour s'adapter aux écrans d'ordinateurs et non aux téléphones. Une fois encore, il a fallu négocier le contenu pour savoir ce qui convenait à ces appareils en fonction de leurs caractéristiques [15].

1.1 Énoncé du problème

Les ressources sur le Web sont accessibles via des identificateurs uniformes de ressources (URI). Une ressource peut avoir différentes alternatives que nous appelons variantes comme dans [7]. Pour demander une variante spécifique, un utilisateur utilise la NC. Ainsi, la NC est le mécanisme permettant de choisir la meilleure variante parmi un ensemble d'alternatives d'une ressource disponible sur le Web sous

un URI. La demande envoyée par un client contient un ensemble de contraintes qui permettent au serveur de fournir une réponse adéquate. Le traitement de la demande par le serveur peut varier en fonction du style et de la dimension de la NC utilisée, ainsi que de la technique de transmission des contraintes et du protocole employé.

1.2 Portée de l'étude

Nous avons utilisé la stratégie suivante pour identifier les caractéristiques de la NC ainsi que les contributions disponibles faisant l'objet de notre classification dans notre état de l'art. Nous avons utilisé le mot clé "content negotiation" pour rechercher des articles dans Google Scholar [1], The DBLP Computer Science Bibliography [2] et Semantic Scholar [3]. Nous avons choisi des articles, des rapports et des études empiriques pertinents parmi les ressources collectées en nous basant sur la compatibilité du problème étudié avec notre énoncé de problème présenté dans la section 1.1. En outre, nous avons suivi les citations et les références afin de considérer les ressources supplémentaires qui citent ou sont citées par les ressources pertinentes ci-dessus. Sept contributions ont été retenues pour être classées dans cet article. Une version à jour du classement est disponible dans la ressource *CNTF* sur laquelle nous travaillons [4].

1.3 Problèmes et études connexes

Des efforts ont été faits pour rassembler les techniques et les caractéristiques disponibles de la NC. Toutefois, à notre connaissance, cela n'a pris la forme que d'une section de travaux connexes dans les articles traitant de la NC, comme [23, Section 5] ou de la documentation de pages Web, comme dans les Mozilla Developer Network Web Docs [5]. Ces efforts ne contiennent que les dimensions générales de la NC. Par exemple ils ne mentionnent pas la dimension de capacité que nous aborderons plus tard, ou ne traitent que certains styles de la NC.
Un autre problème connexe est la recherche d'informations personnalisée, qui vise à aider les utilisateurs à trouver des informations parmi la myriade de ressources disponibles sur le Web. Cependant, contrairement au domaine traditionnel de la recherche d'information, elle prend en compte les préférences de l'utilisateur et l'historique de ses interactions avec le système dans le but principal d'augmenter la satisfaction de l'utilisateur. Une enquête a été menée et a proposé une classification de ces systèmes [11].

1.4 Contribution de l'article

La NC a donné lieu à des résultats de recherche fructueux ces dernières années. Cependant, il n'existe pas d'état de l'art approfondi sur ce sujet de recherche. À notre connaissance, nous fournissons le premier état de l'art des approches et des caractéristiques de la NC existantes. Nous identifions les caractéristiques de chaque contribution et les catégorisons afin de classer les contributions existantes sur la NC. Sur la base de cet état de l'art, nous proposons des cas d'utilisation qui ne sont pas encore bien traités et suggérons des directions pour les travaux futurs.

1.5 Structure de l'article

Le reste de cet article est organisé comme suit. La section 2 présente la terminologie de la NC. La section 3 passe en revue les caractéristiques de la NC et les regroupe en catégories, ce qui nous permet dans la section 4 de fournir une analyse comparative d'une liste de contributions dans ce domaine. Enfin, nous concluons avec la présentation de deux cas d'utilisation que la NC, dans son état actuel, n'est pas en mesure de traiter pleinement. Nous suggérons une orientation future dans la section 5.

2 Terminologie

Cette section présente la terminologie de base de la NC.

Agent Utilisateur (User-Agent) : est un navigateur ou une application mobile par lequel l'utilisateur communique avec le serveur [7].

Client : est le code d'application qui veut accéder aux ressources du serveur et établit des connexions afin d'envoyer des requêtes [7].

Serveur : est un programme d'application qui accepte les connexions afin de répondre aux demandes en renvoyant des réponses. Un programme donné peut être capable d'être à la fois un client et un serveur. Notre utilisation de ces termes fait uniquement référence au rôle joué par le programme pour une connexion particulière, plutôt qu'aux capacités du programme en général. De même, tout serveur peut agir en tant que serveur d'origine, proxy, passerelle ou tunnel, changeant de comportement en fonction de la nature de chaque requête [7].

Ressource : est un objet ou service de données réseau qui peut être identifié par une URI [7].

Variante : est l'une des représentations associées à une ressource à un instant donné [7].

Capacité : est un attribut d'un expéditeur ou d'un destinataire (souvent le destinataire) qui indique une capacité à générer ou à traiter un type particulier de contenu de message [16].

Ressource négociable : est une ressource de données à laquelle est associé plusieurs représentations (variantes). La sélection d'une variante appropriée pour la transmission dans un message est accomplie par la négociation du contenu entre l'expéditeur et le destinataire [16].

Style de négociation : est la manière dont la négociation est menée et de la partie de la NC qui choisit la variante à sélectionner.

Dimension de négociation : également appelée *contraintes* ou *préférences :* indiquent les contraintes prises en compte lors de la sélection de la meilleure représentation.

1. https://scholar.google.com/
2. https://dblp.org/
3. https://www.semanticscholar.org/
4. https://w3id.org/cntf/classification
5. https://developer.mozilla.org/en-US/docs/Web/HTTP/Content_negotiation

3 Caractéristiques de la NC

Nous divisons les caractéristiques de la NC en style, dimension et moyen de transmission des contraintes.

3.1 Dimension de négociation du contenu

Les dimensions de la NC sont les principaux facteurs de différenciation entre un ensemble de variantes. Nous disons que deux représentations alternatives varient en fonction d'une ou plusieurs dimensions de la NC [6].

Type de média : Si un utilisateur a une préférence ou que le client ne peut traiter qu'un type de média spécifique [7], il peut exprimer cela avec HTTP à l'aide de l'en-tête *Accept*.

Langue : Si un utilisateur a une préférence pour une langue particulière, il peut l'exprimer avec HTTP à l'aide de l'en-tête *Accept-Language*.

Encodage : Si un client souhaite que la réponse soit codée à l'aide d'un algorithme de compression particulier. Avec HTTP, il peut l'exprimer à l'aide de l'en-tête *Accept-Encoding*. Les valeurs acceptées sont les suivantes : *gzip*, *compress*, *deflate*, *br*, *identity* ou *.

Char-Set : Avant que *UTF-8* ne soit largement pris en charge, un client pouvait négocier l'encodage des caractères qu'il supportait. Avec HTTP, on l'exprime à l'aide de l'en-tête *Accept-Charset*.

Capacité : En général, un attribut qui définit les capacités et les préférences du matériel ou du logiciel d'un récepteur [8]. Un exemple est la possibilité de traiter les couleurs, les niveaux de gris ou simplement le noir et blanc.

Version : Une ressource pouvant avoir différentes représentations de l'état par lequel elle est passée dans le temps, [6] a introduit un cadre pour négocier les états de la ressource par le biais de la négociation par date, en ajoutant l'en-tête *Accept-Datetime*.

Systèmes de référence de coordonnés : Les objets sur la terre peuvent être localisés à l'aide d'un système de référence de coordonnées. Différents utilisateurs peuvent être intéressés par une représentation de ces objets dans différents systèmes : un système local, régional ou mondial, par exemple WGS84, ETRS89 ou Lambert93.

Autorisation : Les données sur le Web, comme nos comportements dans le monde réel, pourraient être soumises à une réglementation. Les ressources pourraient être limitées, par exemple, pour des raisons de confidentialité. Les utilisateurs peuvent vouloir demander des données uniquement si elles sont conformes à certaines règles et s'ils disposent des autorisations appropriées.

Vocabulaire : Resource Description Framework (RDF) est un langage général permettant de représenter les informations sur le Web. Ces informations sur le Web peuvent être décrites à l'aide de différents vocabulaires. Les utilisateurs peuvent vouloir demander des représentations décrites à l'aide d'un certain vocabulaire, par exemple des données sur une personne décrites par FOAF ou vCard.

Profil OWL : Un profil OWL est une version réduite de *OWL full* qui échange une certaine puissance d'expression contre une efficacité de raisonnement. Pour cette raison, les utilisateurs peuvent souhaiter demander une ontologie dans un profil OWL spécifique tel que *OWL 2 EL*.

Régime d'inférence : Diverses normes du W3C, dont RDF et OWL, fournissent des interprétations sémantiques pour les graphes RDF qui permettent de déduire des déclarations RDF supplémentaires à partir d'assertions explicites. Un utilisateur peut vouloir demander des données conformes à une forme e.g. SHACL/ShEx mais après avoir utilisé un régime d'inférence spécifique tel que RDF Schema, pour valider également les triplets inférés.

Résumé : L'abondance et la longueur des représentations sur le Web rendent leur utilisation de plus en plus difficile pour les humains. Un utilisateur peut vouloir demander un résumé d'une représentation présentant des caractéristiques spécifiques au lieu de recevoir la représentation originale complète.

Mise en page : La mise en page en général fait référence à la manière dont le texte, les images, etc. sont organisés. Un utilisateur peut vouloir demander une représentation qui respecte une mise en page donnée, comme la version minifiée d'un fichier JS, ou le style de citation d'une bibliographie.

Exactitude : La mesure est un processus qui utilise des nombres pour décrire une quantité physique. Les unités de mesure fournissent des normes afin que les nombres dans nos mesures se réfèrent à la même chose. Un utilisateur peut vouloir demander une représentation avec une précision ou une unité de mesure particulière.

Profil : Une définition d'un profil est "une description des contraintes structurelles et/ou sémantiques sur les représentations des ressources qui s'appliquent en plus des contraintes intrinsèquement indiquées par leur type de média" [24]. Un utilisateur peut souhaiter recevoir une représentation uniquement si elle est conforme à un profil particulier, tel que le profil d'application DCAT [23].

3.2 Style de la NC

Dans notre étude, nous avons identifié six styles de la NC : *proactif*, *réactif*, *transparent*, *actif*, *conditionnel*, *adaptatif*. Chacun d'entre eux présente des compromis en termes d'applicabilité et d'aspect pratique [9]. Les sous-sections suivantes présentent une brève description de chaque style, ainsi que certains avantages et inconvénients génériques associés à son utilisation.

3.2.1 Proactive

On parle de négociation proactive (aussi appelée négociation pilotée par le serveur) lorsque les préférences de la NC

6. Notez que cette liste n'est pas exhaustive et que les dimensions ne sont pas orthogonales

7. Liste des types de média : `https://www.iana.org/assignments/media-types/media-types.xhtml`

8. Un exemple de fichier Nokia CC/PP `http://nds1.nds.nokia.com/uaprof/N6230ir200.xml`

9. Il est important de mentionner que les styles de NC ne sont pas mutuellement exclusifs.

sont envoyées par le client dans une requête afin d'encourager un algorithme situé sur le serveur à sélectionner la représentation préférée. La sélection est basée sur les représentations disponibles pour une réponse (les dimensions selon lesquelles elle peut varier, comme la langue, le codage du contenu, etc.) par rapport aux diverses informations fournies dans la requête [8].

Avantages

— Le serveur évite les allers-retours car le client envoie les préférences au serveur qui fait sa meilleure estimation et l'envoie avec la réponse.
— Le serveur n'a pas besoin de décrire l'algorithme de sélection au client pour faire un choix.

Désavantages

— Il est impossible pour le serveur de déterminer avec précision ce qui serait "le mieux" pour un utilisateur donné, car cela nécessiterait une connaissance complète des capacités de l'agent utilisateur et de l'utilisation prévue de la réponse.
— Le fait que l'agent utilisateur doit décrire ses capacités dans chaque requête est un risque potentiel pour la vie privée de l'utilisateur.
— Il complique la mise en œuvre d'un serveur d'origine et des algorithmes de sélection de variantes.
— Il limite la réutilisation des réponses pour la mise en cache partagée.

3.2.2 Réactive

Si le serveur reçoit une demande ambiguë, il envoie une liste des différentes variantes dont il dispose. L'agent utilisateur peut faire le choix s'il a une connaissance suffisante des préférences de l'utilisateur final. Sinon, il affiche la liste des liens pour que l'utilisateur final fasse son choix [8].

Avantages

— Le serveur est incapable de déterminer la capacité de l'agent utilisateur via les en-têtes, ce qui implique plus de confidentialité.
— Le cache peut être utilisé pour réduire la surcharge du réseau.

Désavantages

— Latence due à l'aller-retour pour sélectionner la représentation.

3.2.3 Transparente

La NC transparente est appelée ainsi parce qu'elle rend visibles aux parties intermédiaires (entre le serveur d'origine et l'agent utilisateur) toutes les variantes qui existent au sein du serveur d'origine et leur donne la possibilité de choisir la meilleure représentation en leur nom. La NC transparente est une combinaison de la NC proactive et réactive. Dans la NC réactive, lorsqu'un cache est fourni sous la forme d'une liste de représentations disponibles de la réponse et que la partie intermédiaire (un proxy cache) a bien compris les dimensions de la variance, alors la partie intermédiaire devient capable d'effectuer une NC proactive au nom du serveur d'origine pour les demandes ultérieures sur cette ressource [12, 21, 8].

Avantages

— La réduction du temps de réponse et de la consommation de bande passante en raison de la distribution du travail de négociation qui serait autrement requis du serveur d'origine.
— Le gain du délai de la deuxième demande de négociation réactive lorsque la partie intermédiaire utilise le cache pour pouvoir deviner correctement la bonne réponse.

Désavantages

— Ce style suppose une mise en cache maximale des ressources, ce qui, en pratique, n'est vrai que pour le contenu statique et non crypté.

3.2.4 Conditionnelle

La réponse qu'un serveur donne à une demande dans un style de la NC conditionnelle consiste en un corps composé de plusieurs parties séparées par des limites. Les parties sont rendues de manière sélective en fonction des paramètres de l'agent utilisateur. Cela peut prendre la forme de parties contenant différentes variantes de la ressource, par exemple avec des types de médias distincts [10], ou de parties contenant des portions d'une représentation, par exemple certaines pages d'un document PDF [9, 8] [11].

Avantages

— Réduire le nombre de requêtes émises à une seule qui obtient une réponse composée de plusieurs parties, chaque partie contenant une variante.
— Sélectionner uniquement une partie d'une représentation.

Désavantages

— Ne passe pas à l'échelle si le nombre de variantes ou la taille des variantes sont importants.

3.2.5 Active

Le serveur dans la NC active répond avec une réponse qui contient un script. Le script effectue des demandes supplémentaires (plus spécifiques) en fonction des caractéristiques de l'agent utilisateur [8].

Avantages

— La réduction de l'interaction de l'utilisateur en automatisant l'émission de demandes supplémentaires.
— La fourniture d'une représentation personnalisée qui correspond aux capacités et des besoins de l'agent utilisateur.

Désavantages

— Le besoin d'envoyer plusieurs requêtes pour construire la représentation finale.
— L'introduction de menaces potentielles dues à l'exécution de scripts, par exemple, un attaquant de type man-in-the-middle peut intercepter ou réécrire la réponse pour y inclure du code malveillant.

10. https://docs.marklogic.com/9.0/guide/rest-dev/bulk

11. Dans le cas de parties contenant différentes variantes, ce style est différent du style réactif : dans le premier cas, nous envoyons les variantes réelles séparées par des frontières, alors que dans le second, seuls les URI des variantes disponibles sont envoyés, et des requêtes supplémentaires sont nécessaires pour les récupérer

— L'interdiction du contenu actif par défaut dans les versions les plus récentes des navigateurs en raison des vulnérabilités mentionnées ci-dessus.

3.2.6 Adaptative

Le serveur dans la NC adaptative répond avec une représentation qui a subi un processus d'adaptation. L'adaptation peut être effectuée en interne par le serveur ou en utilisant un autre service pour effectuer cette tâche, par exemple en effectuant une transformation qui nécessite beaucoup de puissance de traitement [19]. L'adaptation est réussie si la représentation finale livrée est plus compréhensible pour l'utilisateur en fonction de son contexte.

Avantages

— Augmenter le taux de satisfaction des contraintes.

Désavantages

— Le processus de pré-adaptation du contenu ne passe pas à l'échelle.

3.3 Transmission de contraintes dans la NC

Dans un processus de la NC, le client doit transmettre au serveur la dimension de négociation ainsi que sa valeur à prendre en compte pour le processus de sélection de la meilleure variante à fournir. Deux techniques principales ont émergé et sont largement utilisées pour effectuer la transmission des contraintes : L'approche *HTTP headers* et l'approche *URL*. Les sous-sections suivantes décrivent chacune de ces deux techniques.

3.3.1 En-têtes HTTP

Les en-têtes HTTP sont des éléments essentiels du protocole HTTP qui permettent la transmission d'informations supplémentaires par le client (en-têtes de requête) et le serveur (en-têtes de réponse). Cette section montrera comment les en-têtes HTTP sont utilisées pour transmettre des contraintes. Il est important de préciser que le protocole CoAP (Constrained Application Protocol) implémente également la NC, mais uniquement pour les type de média via l'option *accept* [22, Section 5.10.4].
La ressource CNTF contient plus de détails [12].

Accept : en-tête de requête qui indique le type de média que l'utilisateur préfère. Le serveur sélectionne une variante et informe le client de son choix avec les en-têtes de réponse *Content-Type* [8, Section 5.3.2].

Accept-Language : une en-tête de requête qui indique la langue exprimée par le langage naturel et la locale que l'utilisateur préfère. Le serveur sélectionne une variante et informe le client de son choix avec les en-têtes de réponse *Content-Language* [8, Section 5.3.5].

Accept-encoding : une en-tête de requête qui indique l'encodage. Typiquement un algorithme de compression que le client préfère. Le serveur sélectionne une proposition et informe le client de son choix avec les en-têtes de réponse *Content-Encoding* [8, Section 5.3.4].

12. `https://w3id.org/cntf`

Accept-Crs : une en-tête de requête qui indique le CRS que l'utilisateur préfère. Le serveur sélectionne une proposition et informe le client de son choix avec les en-têtes de réponse *Content-Crs* [13].

Accept-Presentation : une en-tête de requête qui indique la présentation que l'utilisateur préfère. Le serveur sélectionne une proposition et informe le client de son choix avec les en-têtes de réponse *Content-Presentation* [18].

3.3.2 Basée sur l'URL

Les URI ne fournissent pas seulement un moyen simple et extensible d'identifier des ressources sur le Web, mais ils pourraient aussi être utilisés pour transmettre des contraintes afin de guider la sélection d'une variante préférée. Dans cette section, nous présentons trois façons de les utiliser.

Archival Resource Key (ARK) : L'ARK est un schéma d'identification pour un identifiant persistant des objets d'information [17]. En utilisant la partie facultative "Qualifier", il est possible de créer une sorte de point d'entrée de service qui permet à un ARK de prendre en charge l'accès aux variantes (versions, langues, formats) des composants en utilisant le caractère "." (point) après la partie Nom d'un ARK. Par exemple, l'URI `https://api.istex.fr/ark:/67375/6GQ-MLC8GRWC-5` liste toutes les variantes et la variante PDF se trouve à l'URI `https://api.istex.fr/ark:/67375/6GQ-MLC8GRWC-5/fulltext.pdf`.

Extension de l'URL (suffixe de correspondance de motifs) : cette approche est similaire à l'approche ARK, qui consiste à utiliser l'URI avec une extension principalement pour demander le type de média des points de terminaison de l'API. L'URI `http://myapi.example.com/account/123.json` fournirait une représentation du compte 123 dans un format json.

Query String Arguments (QSA) : Une chaîne de requête est une partie d'une URL qui attribue des valeurs à des paramètres spécifiques. Les QSA sont généralement utilisées pour fournir des informations supplémentaires à un serveur. L'une d'entre elles concerne les préférences pour sélectionner une variante appropriée [14].

4 Classification de certaines contributions existantes

Le tableau 1 présente notre effort pour rassembler quelques contributions connues dans la littérature ayant en commun leur utilisation de la NC. Chaque contribution est présentée avec sa référence, sa date de publication ainsi que ses caractéristiques respectives selon notre catégorisation présentée précédemment.
Composite Capabilities / Preferences Profile (CC/PP) et WAP UAProf sont des descriptions des capacités de péri-

13. `https://github.com/opengeospatial/conneg-by-crs/`
14. e.g. `http://linked.data.gov.au/dataset/gnaf/address/GAACT714845933?_view=ISO19160&_format=text/turtle`

Ref	Date	Style	Dimension	Transmission	Protocole
[4]	2001	Proactive	Capacité	en-tête	HTTP
[20]	2003	Adaptative	Capacité	QSA	HTTP
[25]	2018	Réactive	Type de média	ARK	HTTP
[14]	2018	Réactive	Version	en-tête	HTTP
[18]	2018	Proactive, Adaptative	Présentation	en-tête	HTTP, CoAP
[1]	2018	Conditionnelle	Media type	en-tête	HTTP
[2]	2022		Type de média	en-tête	CoAP

TABLE 1 – Contributions utilisant la NC (triées par date de publication) et leurs caractéristiques (Blanc : non connu).

phériques et des préférences des utilisateurs. [4] décrit une implémentation de la NC HTTP qui les utilise pour fournir la meilleure variante. Le style proactif de la NC est utilisé et la dimension est *capacité*. Les en-têtes sont utilisés pour transmettre les contraintes et le protocole est HTTP.

L'article [20] présente un moteur de décision capable de déterminer les décisions d'adaptation optimales à partir de l'interpolation d'informations contextuelles situationnelles, par exemple la capacité du dispositif. HTTP a été utilisé en intégrant le userid dans l'URL pour identifier l'utilisateur final.

Istex est une archive scientifique française [25]. Les utilisateurs ont la possibilité d'avoir la représentation dans plusieurs formats et pour cela, comme mentionné sur le site, un ARK est utilisé. Si on demande la ressource sans spécifier le type de média, on reçoit un fichier json décrivant les variantes existantes et on peut donc considérer qu'il s'agit d'une négociation réactive. La dimension est le type de média et la transmission dans ARK en utilisant le protocole HTTP.

Les archives sur le Web jouent un rôle majeur en fournissant une image reflétant l'état du Web à un moment donné. La contribution [6] implique la demande du client dans le processus d'agrégation Memento au delà de la spécification de l'URI de la ressource originale et d'une date comme décrit dans [14] en utilisant l'en-tête *prefer*. Par conséquent, la dimension de la NC utilisée est la version, les en-têtes pour le transport des contraintes et HTTP comme protocole. De plus, comme l'agrégateur renvoie un ensemble d'archives et que le client peut potentiellement manipuler la réponse pour émettre une autre requête, nous jugeons que le style de NC est réactif.

Un scénario décrit dans la contribution [18, Section 3.2] est qu'un client demande à un serveur la représentation d'une ressource et souhaite que la réponse soit encodée selon une présentation RDF spécifique, ce qui est fait en incluant des métadonnées dans l'en-tête de la requête. Et parce que cette contribution est principalement destinée aux dispositifs contraints, le style proactif + adaptatif est préféré. L'auteur précise que, même si cette méthode a été implémentée en HTTP, elle pourrait facilement être définie comme des options CoAP équivalentes.

En utilisant le serveur MarkLogic [1], un développeur d'applications REST utilise le style de la NC conditionnelle. Un client recevrait une réponse avec un corps contenant plusieurs parties séparées par un délimiteur à sélectionner. Cette méthode est principalement utilisée pour sélectionner la dimension du type de média et utilise HTTP et les en-têtes pour transmettre cette contrainte. En outre, la spécification de l'Open Connectivity Foundation (OCF) [2] comporte aussi les moyens de négociation du type de média à l'aide du protocole CoAP.

5 Conclusions et travaux futurs

La NC est un mécanisme fondamental du Web. Dans cet état de l'art, nous avons présenté une analyse des caractéristiques des scénarios de la NC existants, notamment les styles, les dimensions, les moyens de transmission des contraintes et les protocoles de la NC. Ces caractéristiques peuvent être utilisées pour classer les contributions existantes comme indiqué dans le tableau 1.

Cependant, certains cas d'utilisation n'ont pas encore de solution satisfaisante, par exemple le cas de la négociation de vocabulaire où un utilisateur veut avoir un moyen de demander des représentations utilisant un vocabulaire spécifique. Par exemple, demander que les données du créateur utilisent le vocabulaire FOAF (Friend Of A Friend), Schema.org ou DCMI (Dublin Core Metadata Initiative). Si les graphes de données disponibles sur le serveur utilisent le même type de média, par exemple : *text/turtle*, le client doit interroger manuellement tous les graphes de données pour sélectionner ceux qui utilisent le vocabulaire souhaité. Une autre limitation est mise en évidence avec le cas d'utilisation de la négociation des formes RDF, où un utilisateur a besoin d'une représentation conforme à une forme spécifique, donc même la négociation du vocabulaire n'est pas suffisante car le client devrait valider manuellement tous les graphes de données retournés. Dans ce cas, la négociation peut être rigide dans le cas où l'utilisateur veut que *toutes* les contraintes soient valides, et préfère ne pas avoir de réponse autrement. En revanche, la négociation peut être flexible dans le cas où l'utilisateur accepte de recevoir une représentation même si elle ne satisfait pas toutes les contraintes. Et enfin, ces contraintes de forme peuvent ne pas avoir le même degré d'importance. L'utilisateur peut donc vouloir un moyen d'exprimer cette importance pour chaque contrainte et d'obtenir la représentation qui en tient compte.

Une direction plausible est l'utilisation des langages de validation pour exprimer des contraintes plus fines, par exemple le langage SHACL du Web sémantique. Plus précisément, l'utilisation de l'en-tête récemment intro-

duit *accept-profile* pour demander une variante qui valide un ensemble de contraintes sous la forme de documents SHACL [23].

Références

[1] REST Application Developer's Guide. Technical report, MarkLogic Corporation, May 2019.

[2] OCF Core specification 2.2.5. Technical report, Open Connectivity Foundation, January 2022.

[3] T. Berners-Lee, R. Cailliau, J. Groff, and B. Pollermann. World-Wide Web : The Information Universe. *Electronic Networking : Research, Applications and Policy*, 2(1) :74–82, 1992.

[4] Mark H. Butler. Implementing content negotiation using CC/PP and WAP UAProf. *HP Laboratories Technical Report HPL*, (190), 2001.

[5] N. Choudhury. World Wide Web and Its Journey from Web 1.0 to Web 4.0. *Int. Journal of Comp. Sci. and Information Tech.*, 5(6) :8096–8100, 2014.

[6] Herbert Van de Sompel, Michael L. Nelson, and Robert Sanderson. HTTP framework for time-based access to resource states - memento. RFC 7089, 2013.

[7] R. Fielding, J. Gettys, J. Mogul, H. Nielsen, L. Masinter, P. Leach, and T. Berners-Lee. Hypertext Transfer Protocol - HTTP/1.1. RFC 2616, IETF, 1999.

[8] R. Fielding and J. Reschke. Hypertext Transfer Protocol (HTTP/1.1) : Semantics and Content. RFC 7231, IETF, June 2014.

[9] Roy Fielding, Yves Lafon, and Julian Reschke. Hypertext transfer protocol (HTTP/1.1) : range requests. RFC 7233, IETF, 2014.

[10] R. Fisher, W. Ury, and B. Patton. *Getting to yes : Negotiating agreement without giving in*. Penguin, 2011.

[11] M Rami Ghorab, Dong Zhou, Alexander O'connor, and Vincent Wade. Personalised information retrieval : survey and classification. *User Modeling and User-Adapted Interaction*, 23(4) :381–443, 2013.

[12] K. Holtman and A. Mutz. Transparent Content Negotiation in HTTP. RFC 2295, IETF, March 1998.

[13] I. Jacobs and N. Walsh. Architecture of the World Wide Web, Volume One. W3c recommendation, W3C, December 15 2004.

[14] Mat Kelly, Sawood Alam, Michael L. Nelson, and Michele C. Weigle. Client-Assisted Memento Aggregation Using the Prefer Header. In *Proc. of WADL'18*, Fort Worth, TX, June 2018.

[15] G. Klyne, F. Reynolds, C. Woodrow, H. Ohto, J. Hjelm, M. Butler, and L. Tran. Composite Capability/Preference Profiles (CC/PP) : Structure and Vocabularies 1.0. W3c recommendation, W3C, January 15 2004.

[16] Graham Klyne. Protocol-independent content negotiation framework. Technical report, IETF, 1999.

[17] John A. Kunze and Emmanuelle Bermès. The ARK Identifier Scheme. Internet draft, IETF, 2008.

[18] M. Lefrançois. 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 2018, Santa Barbara, CA, USA, October 15-18, 2018*, pages 43 :1–43 :8. ACM Press, October 2018.

[19] Tayeb Lemlouma and Nabil Layaïda. Universal profiling for content negotiation and adaptation in heterogeneous environments. In *W3C Workshop on Delivery Context*, pages 4–5, 2002.

[20] W. Lum and F. Lau. User-Centric Content Negotiation for Effective Adaptation Service in Mobile Computing. *IEEE Trans. on Soft. Eng.*, 29(12) :1100–1111, 2003.

[21] Srinivasan Seshan, Mark Stemm, and Randy H Katz. Benefits of transparent content negotiation in http. In *Proceedings of the IEEE Globcom 98 Internet Mini-Conference*. Citeseer, 1998.

[22] Zach Shelby, Klaus Hartke, and Carsten Bormann. The Constrained Application Protocol (CoAP). Technical report, Internet Engineering Task Force, June 2014.

[23] L. Svensson, R. Atkinson, and N. Car. Content Negotiation by Profile. W3C Working Draft, W3C, November 26 2019.

[24] L. Svensson, R. Verborgh, and H. Van de Sompel. Indicating, Discovering, Negotiating, and Writing Profiled Representations. Internet draft, IETF, March 2021.

[25] Pascale Viot and Nicolas Thouvenin. Istex : une nouvelle corde à son ark. *Arabesques*, (88) :18–19, 2018.

Session 4 : Modélisation de connaissances complexes (1)

Apports des méthodologies et techniques de développement logiciel pour l'ingénierie des ontologies: Retour d'expérience des contributions au développement de l'ontologie ETSI SAREF

Maxime Lefrançois[1], Raúl García-Castro[2], María Poveda-Villalón[2], Omar Qawasmeh[3]

[1] Mines Saint-Étienne, Univ Clermont Auvergne, INP Clermont Auvergne, CNRS, UMR 6158 LIMOS, F - 42023 Saint-Étienne France
[2] Open Engineering Group, Universidad Politécnica de Madrid, Spain
[3] Hybrid Intelligence, Capgemini Engineering, 69007 Lyon, France

maxime.lefrancois@emse.fr, rgarcia@fi.upm.es, mpoveda@fi.upm.es, omar.alqawasmeh@capgemini.com

Résumé

L'ingénierie logicielle a toujours eu une grande influence dans l'ingénierie des ontologies. Cet article a pour objectif d'identifier ces influences pour certains des grands thèmes de l'ingénierie logicielle moderne : 1. Ingénierie des besoins ; 2. Modèles de cycle de vie du développement logiciel ; 3. Modularisation ; 4. Patrons ; 5. Environnements de développement ; 6. Nommage des versions ; 7. Contrôle des versions et workflow d'édition ; 8. Automatisation ; 9. Intégration et déploiement continu. Pour chaque thème nous identifions des travaux du domaine de l'ingénierie des ontologies qui s'y rapportent, et apportons un retour d'expérience de notre travail de spécification du cadre de développement et du flux de travail de l'ontologie ETSI Smart Applications REFerence (SAREF), et développement du portail communautaire SAREF.

Mots-clés

Ingénierie logicielle, Agile, DevOps, Ingénierie des Ontologies, SAREF

Abstract

Software engineering has always had a strong influence in ontology engineering. This article aims to identify these influences for some of the major themes of modern software engineering : 1. Requirements engineering ; 2. Software development life cycle models ; 3. Modularization ; 4. Patterns ; 5. Development environments ; 6. Version naming ; 7. Version control and editing workflow ; 8. Automation ; 9. Continuous Integration and Deployment. For each theme we identify work in the field of ontology engineering that relates to it, and provide lessons learned from our work on the specification of the ETSI Smart Applications REFerence ontology (SAREF) development framework and workflow, and development of the Community SAREF Portal for user engagement

Keywords

Software Engineering, Agile, DevOps, Ontology Engineering, SAREF

1 Introduction

L'ontologie Smart Applications REFerence (SAREF) est constituée d'un ensemble modulaire d'ontologies versionnées. SAREF a été promue par la Commission européenne en collaboration avec l'Institut européen des normes de télécommunications (ETSI) dans le but de disposer d'un modèle de données commun pour limiter la fragmentation de l'internet des objets (IoT). L'ontologie SAREF est développée au sein du comité technique SmartM2M de l'ETSI, et est destinée à permettre l'interopérabilité entre les solutions de différents fournisseurs et entre divers secteurs d'activité de l'IoT, contribuant ainsi au développement du marché numérique mondial. Dans cet article nous présentons des résultats du projet ETSI *Specialist Task Force* (STF) 578 récemment terminé, intitulé : "Spécification du cadre de développement et du flux de travail de SAREF, et développement du portail communautaire SAREF pour la participation des utilisateurs". Ce projet avait pour objectif de spécifier le cadre de développement de SAREF et le flux de travail pour accélérer le développement de SAREF et de ses extensions, et développer un logiciel qui sera utilisé pour automatiser la génération du contenu du portail de l'ontologie à partir des sources de SAREF sur la forge ETSI `https://saref.etsi.org/sources/` . La vision finale du projet est de faire en sorte que les industriels utilisateurs de SAREF soient capables d'apporter leur contribution à SAREF et de maintenir SAREF, sans nécessiter de compétences poussées en ingénierie des ontologies ni d'un soutien spécial de l'ETSI, mais juste avec une révision des membres de l'ETSI, et en particulier de SmartM2M.

L'ingénierie logicielle a toujours eu une grande influence dans l'ingénierie des ontologies. Cet article a pour objectif d'identifier ces influences pour certains des grands

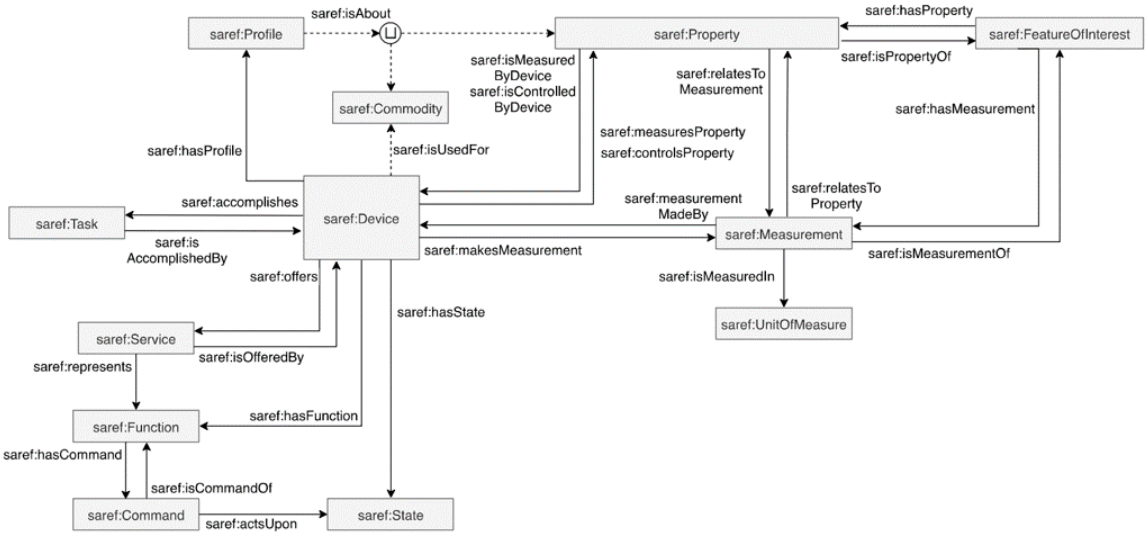


FIGURE 1 – Aperçu des concepts principaux de la version V3.1.1 du module SAREF Core (Source : [23])

thèmes de l'ingénierie logicielle moderne, et d'apporter un retour d'expérience de nos contributions au développement de l'ontologie ETSI SAREF. Nous identifions enfin des difficultés rencontrées ou des directions de travail futures possibles. Nous nous sommes appuyés sur des documents de référence dans le domaine de l'ingénierie logicielle pour sélectionner les thèmes abordés : le *Software Engineering Body of Knowledge (SWEBOK) v3* [36], et le *Systems and Software Engineering Vocabulary (SEVOCAB)* [37]. Ainsi, nous aborderons les thèmes suivants séquentiellement dans cet article : 1. Ingénierie des besoins ; 2. Modèles de cycle de vie du développement logiciel ; 3. Modularisation ; 4. Patrons ; 5. Environnements de développement ; 6. Nommage des versions ; 7. Contrôle des versions et workflow d'édition ; 8. Automatisation ; 9. Intégration et déploiement continu. Il est à noter que de nombreux travaux démontrent les apports des ontologies pour différents sous-domaines de l'ingénierie logicielle, par exemple pour l'élicitation des besoins [40], cependant nous considérons ces travaux hors du cadre de cet article.

2 Ingénierie des besoins

L'ingénierie des besoins en ingénierie logicielle se rapporte à l'élicitation, l'analyse, la spécification, et la validation des besoins logiciels. Dès Grüninger et Fox en 1995 [32], On-To-Knowledge [61] en 2001, eXtreme ontology method [35] en 2002, les méthodologies d'ingénierie des ontologies comportent habituellement une phase de spécification des besoins sous forme de questions de compétences [64], sous une forme textuelle, de logique du premier ordre, ou sous forme d'une requête SPARQL. L'étude des techniques d'élicitation des connaissances pour dériver des questions de compétences a été étudiée par exemple par Rao et. al. [58]. Le concept de *Software Specification Document* a été transposé aux ontologies : le *Ontology Requirement Specification Document* [62]. La thèse récente de Alba Fernández Izquierdo [38] porte sur la spécification et l'évaluation automatique des besoins pour l'ingénierie des ontologies.

Dans le contexte du projet STF 578 de l'ETSI, nous recommandons la présence d'un document spécifiant les besoins pour tout projet d'ontologie SAREF [22, Clause 9], sous forme d'un document CSV avec trois colonnes : un identifiant, une catégorie, et un besoin exprimé sous forme d'une affirmation ou d'une question de compétences. Ces besoins sont ensuite évalués avec l'outil Themis [25].

3 Modèles de cycle de vie du développement logiciel

De nombreuses méthodologies d'ingénierie des ontologies ont été proposées au fil du temps, dont METHONTOLOGY [26], On-To-Knowledge [61], DILIGENT [54], le "Ontology Development 101" [50], NeOn [29]. Certaines transposent directement des méthodologies d'ingénierie logicielle, par exemple UPON Lite [10] est basé sur Rational Unified Process. Les récentes méthodologies s'inspirent toutes des principes d'ingénierie logicielle Agile. AMOD [1] et CD-OAM [63] sont basés sur SCRUM, XPOD [59] et eXtreme ontology method [35] sont basés sur eXtreme Programming, Lean Ontology Development (LOD) [9] s'inspire de l'approche Lean, SAMOD [53] se base sur les concepts de stories, itérations, et développement dirigé par les tests. Le développement des ontologies SAREF suit la méthodologie LOT [55], qui adopte une approche de type V-model avec des retours conditionnels à des étapes de développement amont.

Cette liste n'est pas exhaustive, mais démontre que le domaine de l'ingénierie logicielle inspire et influence celui de l'ingénierie des ontologies. Il est donc pertinent de surveiller les évolutions du premier domaine pour continuer à améliorer le second.

4 Modularisation

La modularité est définie comme *le degré selon lequel un système ou un programme informatique est composé d'éléments distincts, de sorte que la modification d'un élément a un impact minimal sur les autres éléments* [37]. Une transposition aux ontologies et un premier algorithme de modularisation ont été proposés dans [31]. Le sujet intéresse la communauté, voir par exemple les séries de workshops WoMO (Workshop on Modular Ontologies, de 2006 à 2013) et WOMoCoE (Workshop on Ontology Modularity, Contextuality, and Evolution, de 2016 à 2020). Beaucoup d'ontologies sont aujourd'hui publiées sous forme de réseau d'ontologies, composé de modules faiblement dépendants qui utilisent le mécanisme d'import de OWL [48, Sec. 3.4]. Une ontologie noyau n'importe aucune autre ontologie du réseau, des ontologies périphériques importent l'ontologie noyau et potentiellement d'autres ontologies périphériques, enfin des modules d'alignement importent au moins une ontologie du réseau et une ontologie externe.

Comme illustré sur la figure 2, la suite d'ontologies ETSI SAREF est composée d'ontologies définissant des patrons génériques comme SAREF4SYST [21], d'une ontologie noyau SAREF Core [23] illustrée dans la figure 1, et de différentes extensions développées pour des domaines verticaux distincts : SAREF4ENER pour l'energie [12], SAREF4ENVI pour l'environement [14], SAREF4BLDG pour les bâtiments intelligents [14], SAREF4CITY pour les villes intelligentes [15], SAREF4INMA pour les l'industrie manufacturière [16], SAREF4AGRI pour l'agriculture [17], SAREF4AUTO pour l'automobile [18], SAREF4EHAW pour la e-santé et le bon vieillissement [19], SAREF4WEAR pour les wearables [20], SAREF4WATR pour la gestion de l'eau [13], SAREF4LIFT pour les ascenseurs intelligents [24].

Deux choix de conception sont récurrents dans ces réseaux d'ontologies : (1) la définition des espaces de noms pour chaque module, et (2) le choix du module dans lequel un terme est défini. Choisir un espace de nom distinct pour chaque module permet d'identifier facilement de quel module un terme est issu. Cela simplifie également la publication des ontologies dans le respect de la bonne pratique qui consiste à rendre une description de chaque terme accessible à son IRI (des IRI de type hash '#' peuvent être utilisées). Cependant, cette approche pose trois problèmes :

Premièrement, il est parfois difficile en tant qu'utilisateur de ces ontologies, de se souvenir quel est l'espace de nom pour chaque concept. Nous avons par exemple une variété de sous-classes de saref:Property réparties dans les espaces de noms des différentes extensions, selon là où on a eu besoin de les définir en premier : saref:Temperature, saref:Humidity, saref:Power, s4ener:Power, s4ener:PowerMax, s4ener:PowerStandardDeviation, s4inma:Size, saref:Light, s4envi:LightProperty, ainsi que les instances de saref:Property suivantes : s4envi:Frequency, s4wear:SoundLevel, s4wear:BatteryRemainingTime, s4watr:Conductivity, s4wear:Temperature. Une approche alternative aurait consisté à utiliser un espace de nom unique et des IRIs de type slash '/', et implémenter des redirections de l'IRI de chaque terme vers le document qui décrit l'ontologie où il est défini.

Deuxièmement, l'expérience montre qu'il peut être pertinent de déplacer un terme d'un module vers un autre. Par exemple SAREF4CITY V1.1.1 a introduit le concept de s4city:FeatureOfInterest, et il a été décidé lors du développement de SAREF Core V3.1.1 que ce concept devait être déplacé dans l'ontologie noyau. Il est donc maintenant identifié par saref:FeatureOfInterest, et les implémentations de SAREF4CITY ont dû être modifiées. Ce problème ne se serait pas posé si une approche basée sur un espace de nom unique et des IRIs de type slash avait été adoptée.

Enfin, ce que montre la liste des classes et instances de saref:Property, c'est qu'au sein même de la communauté des développeurs de SAREF, des choix de modélisation et de nommage sont parfois variés. Il nous apparaîtrait donc important de re-baser le développement de SAREF sur des patrons d'ontologies pour harmoniser son développement.

5 Patrons

En ingénierie logicielle, un patron est défini comme une *spécification abstraite d'une composition d'objets qui fait que toute instance de la composition possède une propriété donnée.* [37]. Le concept a été transposé à l'ingénierie des ontologies avec les *Ontology Design Patterns* [27, 6, 28]. C'est le sujet par exemple de l'état de l'art [7].

Un des livrables du projet STF 556 de l'ETSI est le rapport technique TR 103 549 nommé "Consolidation de SAREF et de sa communauté d'utilisateurs industriels, sur la base de l'expérience du projet EUREKA ITEA 12004 SEAS". Ce rapport identifie les patrons implicitement existant dans SAREF, et qu'il pourrait convenir de formaliser pour aboutir à une version consolidée de l'ontologie SAREF [11]. Par exemple dans la version V2.1.1 de SAREF Core, les fonctions de détection, d'actionnement et de mesure sont des types de fonctions. Habituellement, une fonction (par exemple saref:StartStopFunction) a une ou plusieurs commandes pour la déclencher (par exemple, pour saref:StartStopFuncton, ce devrait être soit une saref:StartCommand soit une saref:StopCommand). Certaines commandes agissent sur certains états (saref:StartStopCommand agit sur un certain saref:StartStopState). Il conviendrait par exemple de s'assurer que toutes les sous-classes de la classe saref:Command soient décrites de la même manière. Par exemple, des sous-classes de saref:Command avaient des instances génériques, associées à aucune réelle action. SAREF avait aussi une commande nommée saref:PauseCommand, qui n'était associée à aucune fonction.

Des patrons peuvent être instanciés avec les éléments pris dans un ou plusieurs dimensions orthogonales. Par exemple, SAREF4ENER définit s4ener:EnergyMax, s4ener:EnergyMin, s4ener:EnergyExpected, s4ener:EnergyStandardDeviation, s4ener:PowerMax, s4ener:PowerMin, s4ener:PowerExpected, s4ener:PowerStandardDeviation. Gérer manuellement l'ajout par exemple d'un

nouveau type d'agrégat *Average* implique de créer de nombreuses propriétés, comme s4ener:EnergyAverage, s4ener:PowerAverage.

Une solution partielle à ce problème consiste à découpler les dimensions. Dans l'exemple ci-dessus : le type de propriété, et le type d'agrégat. Dans les ontologies SEAS [45] nous avons proposé une modélisation qui vise à éviter ces situations, en découplant les dimensions. Une entité d'intérêt est liée à une seule instance de la classe seas:Property par le biais d'une relation qui serait nommée seas:hasElectricConsumption par exemple. Cette instance de propriété peut alors être d'un type générique seas:PowerProperty, et des évaluations de cette propriété peuvent être définies et multi-typées. Par exemple seas:Evaluation, seas:MinEvaluation, seas:AverageEvaluation, seas:SumEvaluation.

De plus, vouloir modifier généralement comment sont décrites les sous-classes d'une classe principale comme saref:Property peut s'avérer fastidieux, car chaque instance du patron doit être revue manuellement. Parmi les travaux récents qui permettent d'automatiser la génération des ontologies à partir de patrons et de description des instances, on peut citer le système *Reasonable Ontology Template* OTTR [1] [60], *Generic Ontology Design Patterns* [43, 44], ou *Dead Simple OWL Design Patterns* (DOS-DPs) [52]. OTTR permet de déclarer des patrons d'ontologie, et de générer automatiquement des instances de ces patrons à partir d'un document externe, par exemple un document Excel écrit par les experts de domaine.

L'ontologie SAREF4SYST [21], illustrée sur la figure 3 et inspirée de SEAS, est la première ontologie de patron incorporée à SAREF. Elle définit un modèle d'ontologie qui peut être instancié pour différents domaines. SAREF4SYST définit les systèmes, les connexions entre les systèmes et les points de connexion auxquels les systèmes peuvent être connectés. Ces concepts de base peuvent être utilisés de manière générique pour définir la topologie des entités d'intérêt, et peuvent être spécialisés pour de multiples domaines. Par exemple, pour décrire des zones à l'intérieur d'un bâtiment (systèmes), qui partagent une frontière (connexions). Les propriétés des systèmes sont généralement des variables d'état (par exemple, la population des agents, la température), tandis que les propriétés des connexions sont généralement des flux (par exemple, le flux de chaleur). SAREF4SYST a deux objectifs principaux : d'une part, étendre SAREF avec la capacité de représenter la topologie générale des systèmes et comment ils sont connectés ou interagissent et, d'autre part, illustrer comment les patrons d'ontologie peuvent aider à assurer une structure homogène de l'ontologie SAREF globale et accélérer le développement d'extensions.

1. https://ottr.xyz/

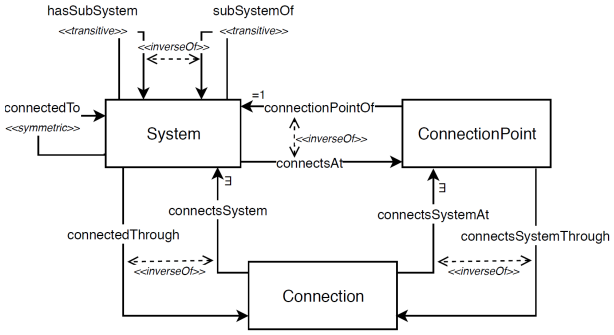


FIGURE 3 – Aperçu du patron d'ontologie SAREF4SYST (Source : [21])

6 Environnement de développement

Différents logiciels spécialisés existent pour l'édition des ontologies [2], Stanford Protégé [49] étant probablement le plus connu et utilisé. Des logiciels professionnels existent comme TopBraid Enterprise Vocabulary Net [2] par exemple, qui est une extension de l'environnement de développement intégré Apache Eclipse.

Pour l'édition du code source avec la syntaxe Turtle 1.1 par exemple, des plugins commencent à fleurir pour la plupart des éditeurs de texte multi-langage. Par exemple le plugin *Linked Data syntaxes* [3] pour Sublime Text [4], ou l'extension

2. https://www.topquadrant.com/the-topbraid-evn-ontology-editor/
3. https://github.com/blake-regalia/linked-data.syntaxes
4. https://www.sublimetext.com/

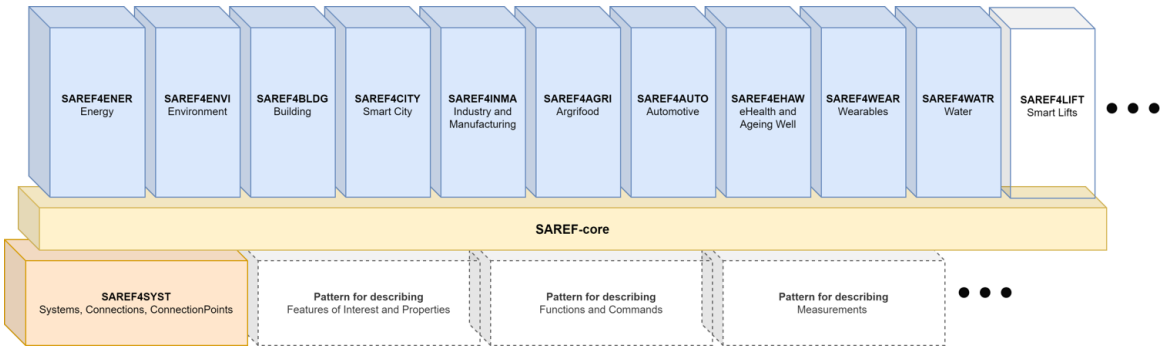


FIGURE 2 – L'ontologie SAREF et ses différents modules

Stardog RDF Grammars [5] pour Visual Studio Code [6]. Ces extensions permettent la coloration syntaxique des fichiers RDF, et ainsi d'identifier rapidement des erreurs de syntaxe. Des fonctionnalités supplémentaires peuvent être attendues d'un environnement de développement intégré d'ontologies. Nous avons *forké* par exemple le plugin *Linked Data syntaxes* pour implémenter l'exécution de règles SPARQL-Generate [46] avec la combinaison de touche CTRL+B [7]. Un récent projet de nos étudiants [8] consistait à implémenter une fonctionnalité de navigation dans des projets contenant de nombreux fichiers RDF : un contrôle-clic sur un terme RDF permet de naviger vers là où est défini le terme en question : soit un fichier local, soit une URL dans le navigateur. Lorsque l'on écrit une IRI préfixée, le système peut également vérifier que le préfixe est déclaré, et éventuellement insérer automatiquement le nouveau préfixe dans l'entête si il est connu de la base de données `prefix.cc` [9].

Le *Linting*, dont le nom vient d'une commande UNIX de pré-processeur pour le langage C, est une approche qui consiste à analyser statiquement le code source d'un logiciel pour détecter des erreurs, bugs, ou erreurs de style. Développer un *linter* pour supporter le processus d'édition des ontologies permettrait de limiter la difficulté d'éditer des ontologies de qualité. Par exemple avec Jena Eyeball, [10], ou RDFLint [11]. Ce dernier est intégré dans l'extension *RDF language support via rdflint* [12] de Visual Studio Code, et permet entre autre d'exécuter des requêtes SPARQL, de valider des contraintes SHACL, ou de valider que les littéraux sont bien formés. Bien que le développement de tels outils soit difficile à valoriser d'un point de vue de la recherche, ils nous semblent extrêmement importants pour contribuer à abaisser le niveau de compétence nécessaire pour développer des ontologies d'une bonne qualité.

7 Nommage des versions

Les ontologie sont amenées à évoluer, pendant le travail de développement potentiellement collaboratif, mais également après une première publication si des évolutions sont nécessaires. Différentes pratiques de nommage des versions existent pour les logiciels, les plus connus étant le *Semantic Versioning* [13] et le *Calendar Versioning* [14]. Une transposition de semver aux ontologies a été proposée [65], et est assez communément utilisé aujourd'hui.

Le langage OWL définit deux types d'identifiants pour les ontologies : un identifiant de série d'ontologie (l'instance de owl:Ontology), et l'identifiant de version d'ontologie (l'objet de la métadonnée owl:versionIRI) [48, Sec. 3.3]. Les utilisateurs peuvent alors choisir d'importer une ontologie par son identifiant de série et ainsi suivre les évolutions de l'ontologie, ou son identifiant de version et s'assurer que rien ne cassera en cas d'évolution non rétrocompatible.

Une erreur de conception commune même à certaines ontologies du W3C consiste à inclure l'identifiant de version, ou la date de publication dans l'identifiant de série d'ontologie. Par exemple, QUDT V1.1 avait l'identifiant `http://qudt.org/1.1/schema/qudt#` (corrigé depuis la version 2). Les identifiants de RDF, RDFS, et OWL, contiennent respectivement `1999/02/22`, `2000/01`, et `2002/07`. Il n'est dans ce cas possible de conserver l'identifiant pour la version suivante de l'ontologie sans créer une incohérence. RDFS 1.1 publié en 2014 conserve l'année 2000 dans son identifiant.

L'évolution d'une ontologie peut avoir différents impacts sur les artefacts (ontologies, bases de connaissances, logiciels) qui l'utilisent. Dans sa thèse, Omar Alqawasmeh [5, 57] étudie ces différents problèmes et propose des contre-mesures. Une recommandation adoptée dans le travail de développement des ontologies SAREF est par exemple de s'assurer que l'on importe des ontologies par leur IRI de version, et non pas par l'identifiant de série d'ontologie. Par exemple, SAREF4LIFT V1.1.1 importe SAREF Core V3.1.1, SAREF4SYST V1.1.2, et SAREF4BLDG V1.1.2.

Avec SAREF, nous allons plus loin dans l'adoption du *Semantic Versioning*. Chaque module de l'ontologie possède une version distincte, composée de trois numéros : un *MAJOR*, un *Minor*, et un *patch*. L'incrémentation du *MAJOR* indique une coupure de la rétrocompatibilité. L'incrémentation du *Minor* indique l'ajout de fonctionnalités. L'incrémentation du *patch* indique la correction d'un bug. On pourrait donc importer un module SAREF, par exemple SAREF Core, avec différentes IRI, exprimant différents choix : 1. `https://saref.etsi.org/core/` redirigera vers la dernière version Vx.y.z 2. `https://saref.etsi.org/core/v3` redirigera vers la dernière version V3.y.z 3. `https://saref.etsi.org/core/v3.1` redirigera vers la dernière version V3.1.z . On pourrait également proposer un schéma général d'IRI, s'appuyant sur l'expression d'une spécification de version comme le *Maven Dependency Version Range* [15] pour Java, ou PEP440 [16] pour Python.

8 Contrôle des versions et workflow d'édition

Le logiciel de gestion de versions distribuées git s'est rapidement imposé dans l'édition collaborative de logiciels, et les plateformes de type GitHub ou Gitlab ont démocratisé différents flux opérationnels (*workflows*) basés sur

5. https://marketplace.visualstudio.com/items?itemName=stardog-union.stardog-rdf-grammars
6. https://visualstudio.microsoft.com/
7. https://w3id.org/sparql-generate/sublime.html
8. https://github.com/clement000/linked-data.syntaxes
9. https://prefix.cc/
10. https://jena.apache.org/documentation/archive/eyeball/eyeball-manual.html
11. https://github.com/imas/rdflint
12. https://marketplace.visualstudio.com/items?itemName=takemikami.vscode-rdflint
13. https://semver.org
14. https://calver.org
15. https://cwiki.apache.org/confluence/display/MAVENOLD/Dependency+Mediation+and+Conflict+Resolution
16. https://peps.python.org/pep-0440/#version-specifiers

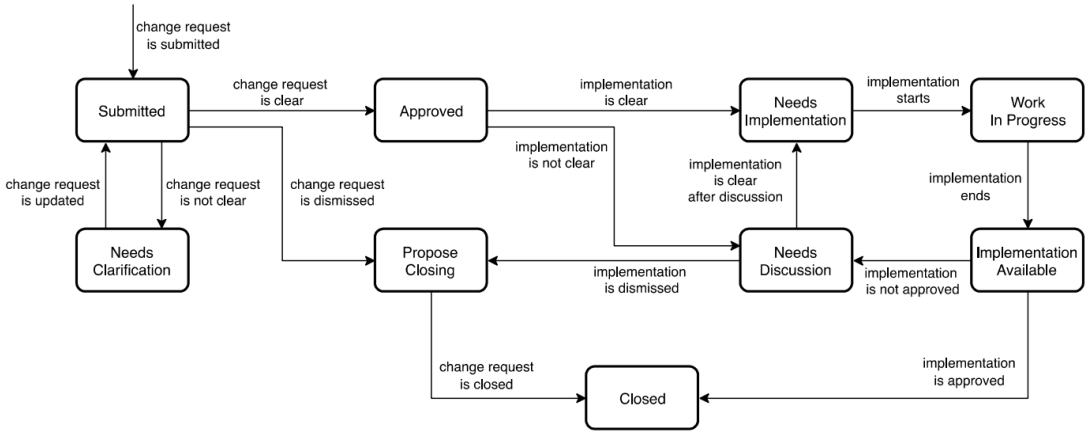


FIGURE 4 – Exemple du flux opérationnel pour l'édition d'une version de SAREF (Source : [22])

les branches (fonctionnalité offerte par git), les forks, tickets, jalons, et requêtes de fusion (*pull request* pour Github, *merge request* pour Gitlab).

Des retours d'expérience de workflows sur Github et identification des meilleures pratiques commencent à être publiés [8, 3]. Pour l'édition des ontologies SAREF, nous utilisons le portail publique ETSI Forge `https://saref.etsi.org/sources/`, et avons publié une spécification technique établissant différents flots de travail pour (1) la création d'une version d'ontologie, (2) le développement d'une version d'ontologie, (3) la publication d'un projet [22, Clauses 6.1, 7.1, 8.1]. Nous utilisons quatre type de branches : des branches `issue-x` pour le travail sur un ticket, des branches `develop-vx.y.z` pour le travail sur une version, des branches `prerelease-vx.y.z` pour le travail de validation final de l'ontologie, et des branches `release-vx.y.z` pour les versions publiées. Des règles de protection sont définies pour interdire aux développeurs d'une ontologie de pousser directement des changements sur les branches `develop-vx.y.z`, ou d'accepter directement des *merge request* sur les branche `prerelease-vx.y.z`. Les tickets sont disponibles à l'adresse `https://labs.etsi.org/rep/groups/saref/-/issues`.

Agec git, une version d'un logiciel, nommé *commit*, contient zéro ou plusieurs *commit* parent, la liste des fichiers modifiés dans ce commit, les versions compressées des nouvelles versions de ces fichiers, et d'autres métadonnées comme la date et l'identifiant de l'auteur du *commit*. Instaurer l'état d'un logiciel pour un commit donné consiste donc à parcourir l'arbre de ses ancêtres, et décompresser pour chaque commit ancêtre les fichiers qui n'ont pas été modifiés ultérieurement.

Git peut se baser sur différents algorithmes de calcul des différences entre fichiers texte (*diff*), défini par la variable `diff.algorithm`. Pour l'édition des ontologies, ceci pose un problème car les éditeurs spécialisés peuvent complètement transformer la sérialisation d'une ontologie, puisque chacun se base sur une librairie de sérialisation différente (Apache RIOT pour Protégé, RDF4J RIO pour Topbraid-Composer). En pratique, cela rend très difficile l'évaluation des changements implémentés lorsqu'il faut valider une Pull Request, puisque tout semble avoir été changé.

Il serait théoriquement possible de modifier l'algorithme qu'utilise git pour la détection de différences entre deux versions, par exemple pour Promptdiff [51], Ecco [30], ou OWLDiff [42]. Cependant il faudrait que le serveur GitHub ou GitLab puisse utiliser le même algorithme pour visualiser le résultat du *diff*. Une approche alternative proposée par exemple dans [33] consiste à utiliser les crochets *hooks* [17], qui s'assurent que le même outil de serialisation est utilisé avant chaque commit.

Deux pratiques principales existent pour identifier des versions de logiciels avec git : les étiquette (*tag*) de version, et les branches de sortie (*release branch*). Le développement des ontologies SAREF utilise cette deuxième approche, qui permet de continuer à faire évoluer la documentation ou les exemples même lorsque l'on fige une ontologie.

9 Automatisation

L'automatisation est un sujet majeur en développement logiciel et a pour objectif d'accélérer la production et la qualité des logiciels, éviter les tâches redondantes, et limiter les mauvaises versions de logiciels. Différentes tâches peuvent être automatisées en ingénierie des ontologies, par exemple avec les outils de *linting* présentés dans la section 6, ou l'interface en ligne de commande disponible dans des frameworks comme Apache Jena [18]. L'outil ROBOT développé par la communauté OBO [39] permet d'exécuter automatiquement un certain nombre de tâches pour convertir, raisonner, importer, extraire des modules, filtrer des axiomes, requêter ou vérifier la bonne exécution de requêtes SPARQL pour évaluer des tests unitaires, générer un rapport d'erreurs, réparer si possible, instancier des patrons, et assembler ces tâches dans des workflows.

Pour les projets de développement d'ontologies qui utilisent

17. `https://git-scm.com/book/fr/v2/Personnalisation-de-Git-Crochets-Git`
18. `https://jena.apache.org/documentation/tools/index.html`

git, il est possible d'extraire automatiquement des informations à injecter dans les métadonnées de l'ontologie. Par exemple les trois commandes suivantes trouvent pour une ontologie `onto.ttl` : 1. la date de premier commit, ce qui peut permettre de renseigner la propriété dct:created, 2. la date de dernière modification (propriété dct:modified), 3. les auteurs classés par ordre décroissant de nombre de modifications (propriété owl:contributor).

```
git log --diff-filter=A --format='%ad' --date=short -- onto.ttl
git log -1 --format='%ad' --date=short -- onto.ttl
git log -- onto.ttl | grep Author | sort | uniq -c | sort -n -r
```

Dans le projet STF 578, nous avons spécifié un ensemble de règles auxquelles un dépôt d'ontologie SAREF doit se conformer dans la spécification technique ETSI TS 103 673 [22, Clause 9], et avons développé l'application SAREF Pipeline qui permet d'évaluer chacune de ces règles avec un niveau d'exigence. voici une liste non exhaustive des points évalués : (a) Structure du répertoire de dépôt, (b) présence d'un fichier de licence défini, (c) spécification des besoins de l'ontologie, (d) présence d'un fichier `/saref4[a-z]{4}.ttl/` bien formé, (e) déclaration de préfixes conformes, (f) présence d'une déclaration d'ontologie, avec une IRI de série et une IRI de version conformes au nommage de la branche git (ex : `develop-v2.1.1`, (g) imports éventuels d'autres ontologies SAREF par leur IRI de version, (h) présence des créateurs et contributeurs, (i) convention de nommage pour les classes, propriétés, instances, (j) présence de métadonnées pour les termes, (k) l'ontologie doit être OWL 2 DL, (l) l'ontologie doit être consistante, (m) chaque classe doit être satisfiable (n) aucun pitfall détecté par OOPS ! [56], (o) présence de tests, (p) présence et qualité des exemples, (q) existence des termes utilisés.

Certains de ces tests utilisent des shapes SHACL [41], d'autres les fonctionnalités de OWLAPI après avoir cloné les dépôts nécessaires. Le dossier des messages de l'application donne une vue globale de toutes les erreurs qui peuvent être identifiées [19]. Cette application peut être utilisée avec une interface graphique (figure 5) ou en ligne de commande (figure 6). Le rapport d'erreur est formatté en markdown, ce qui permet d'ouvrir rapidement un ticket pour traiter le problème collaborativement (figure 7). Finalement, l'application génère différentes sérialisations pour les ontologies et les exemples, et une documentation HTML inspirée de LODE et réécrite avec SPARQL-Generate [46]. Voir par exemple `https://saref.etsi.org/core` ou `https://saref.etsi.org/core/Command`.

Cette application est monolithique et peut difficilement être réutilisée pour d'autres projets d'ingénierie d'ontologie. Nous travaillons sur des améliorations pour les tâches d'ingénierie des ontologies pour les projets ANR HyperAgents (ANR-19-CE23-0030-01) et ANR CoSWoT (ANR-19-CE23-0012-04).

19. `https://labs.etsi.org/rep/saref/saref-pipeline/-/tree/master/src/main/resources/messages`

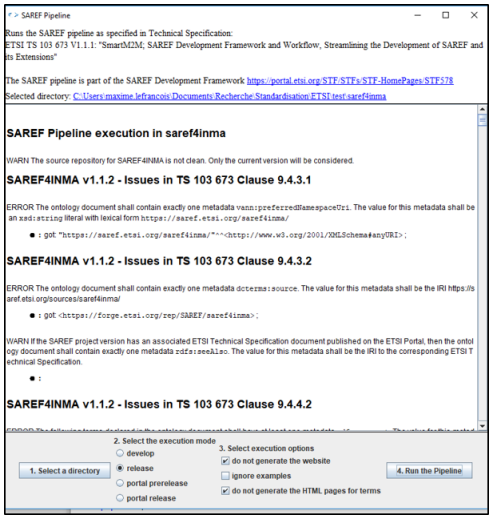


FIGURE 5 – Exécution du pipeline SAREF avec l'interface graphique `https://saref.etsi.org/sources/saref-pipeline/`

```
saref-pipeline/target$ java -jar saref-pipeline.jar
usage: java -jar saref-pipeline.jar <mode> [<options>] [<target>]

Runs the SAREF pipeline as specified in Technical Specification:
  ETSI TS 103 673 V1.1.1: "SmartM2M; SAREF Development Framework and
            Workflow, Streamlining the Development of SAREF and its
            Extensions"

<mode> can take the following values:
  develop   Run the SAREF pipeline in relax mode in the target SAREF
            project
  release   Run the SAREF pipeline in strict mode in the target SAREF
            project
  prerelease-portal    Operate a strict check and generate the portal for
            pre-release or release branches of each source in the
            configuration file `.saref-repositories.yml`
  release-portal    Operate a strict check generate the portal for release
            branches of each source in the configuration file
            `.saref-repositories.yml`
  clean     Remove all files generated by the previous executions
  help      Displays this message

<target> points to the target directory. By default, target is the current
            directory.

<options> can take the following values:
 -e,--no-examples   Do not check examples
 -s,--no-site       Do not generate the static portal
 -t,--no-terms      Do not generate the static portal for terms
```

FIGURE 6 – Exécution du pipeline SAREF en ligne de commande `https://saref.etsi.org/sources/saref-pipeline/`

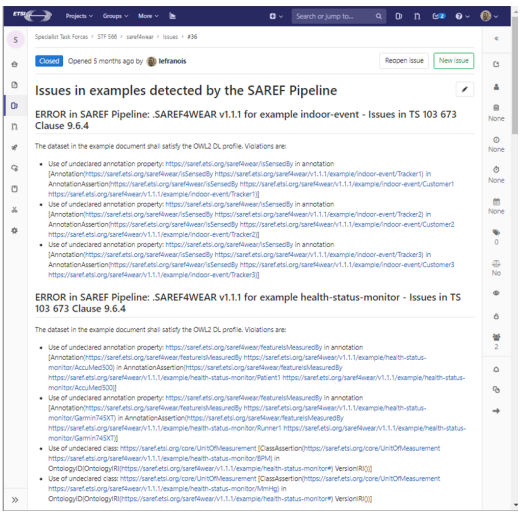


FIGURE 7 – La sortie du pipeline SAREF est formatée en markdown et peut être utilisée pour créer un ticket

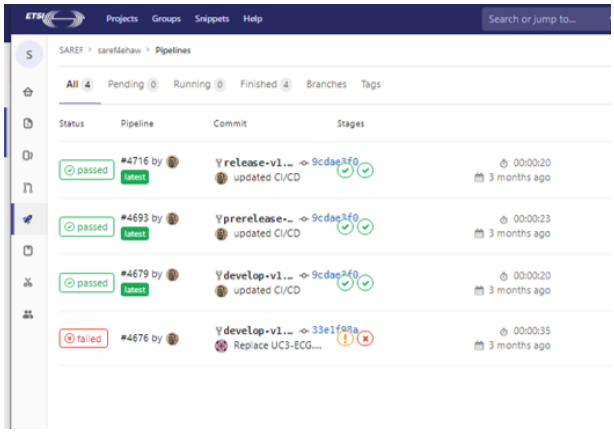


FIGURE 8 – Aperçu des pipeline d'intégration et de déploiement continu : *Snapshot*, *Staging*, *Manual release*. Source : `https://saref.etsi.org/sources/saref4ehaw/-/pipelines`

10 Intégration et deploiement continu

A l'instar des méthodes Agile qui visent à améliorer les collaborations entre les clients d'un projet logiciel et les développeurs, les méthodes DevOps améliorent les collaborations entre les développeurs et les professionnels des opérations informatiques. Jenkins [20], Travis CI [21], Circle CI [22], Gitlab CI/CD [23], Github Actions [24], sont tous des frameworks qui permettent de spécifier des pipelines de tâches qui seront exécutés automatiquement lorsque par exemple un commit est poussé sur le serveur. Avant la démocratisation de ces frameworks, quelques approches préliminaires ont été proposées dans la communauté de l'ingénierie des ontologies en utilisant les applications Github [25]. Par exemple VoCol [34] ou OnToology [4]. *Ontology Development Kit* (ODK) [47] utilise Travis CI pour exécuter des workflows avec ROBOT.

Dans le projet STF 578, nous avons configuré Gitlab CI/CD dans chaque dépôt des ontologies SAREF, pour qu'il exécute le pipeline SAREF différemment selon le type de branche où est poussé un commit (issue, develop, prerelease, release), et pousse finalement automatiquement les fichiers de sortie vers le portail de documentation de SAREF `https://saref.etsi.org/`. La figure 8 illustre l'exécution automatique des pipelines SAREF.

11 Conclusion

Dans cet article nous avons montré que les méthodologies et techniques de développement logiciel ont eu des répercussions importantes en ingénierie de l'ontologie, au moins pour les neuf thématiques identifiées. Nous avons illustré comment le framework de développement et de publication de l'ontologie ETSI SAREF a été spécifié pour tirer partie des dernières méthodologies et techniques disponibles. Nous appliquons et améliorons actuellement ces travaux dans les tâches d'ingénierie des ontologies pour les projets ANR HyperAgents (ANR-19-CE23-0030-01) et ANR CoSWoT (ANR-19-CE23-0012-04)

20. `https://www.jenkins.io/`
21. `https://travis-ci.org/`
22. `https://circleci.com/`
23. `https://docs.gitlab.com/ee/ci/`
24. `https://github.com/features/actions`
25. `https://docs.github.com/en/developers/apps`

Références

[1] Abdelghany Salah Abdelghany, Nagy Ramadan Darwish, and Hesham Ahmed Hefni. An agile methodology for ontology development. *International Journal of Intelligent Engineering and Systems*, 12(2) :170–181, 2019.

[2] Emhimed Salem Alatrish. Comparison of ontology editors. *eRAF Journal on Computing*, 4 :23–38, 2012.

[3] Dean Allemang, Pawel Garbacz, Przemyslaw Gradzki, Elisa Kendall, and Robert Trypuz. An infrastructure for collaborative ontology development, lessons learned from developing the financial industry business ontology (FIBO). In *Formal Ontology in Information Systems*. IOS Press, 2022.

[4] Ahmad Alobaid, Daniel Garijo, María Poveda-Villalón, Idafen Santana-Perez, Alba Fernández-Izquierdo, and Oscar Corcho. Automating ontology engineering support activities with ontoology. *Journal of Web Semantics*, 57 :100472, 2019.

[5] Omar Alqawasmeh. *Towards a collaborative framework for ontology engineering : Impact on ontology evolution and pitfalls in ontology networks and versioned ontologies*. Theses, Université de Lyon, September 2020.

[6] Eva Blomqvist and Kurt Sandkuhl. Patterns in ontology engineering : Classification of ontology patterns. In *International Conference on Enterprise Information System*, pages 413–416, 2005.

[7] Giuseppe Cota, Marlena Daquino, and Gian Luca Pozzato. *Applications and Practices in Ontology Design, Extraction, and Reasoning*, volume 49. IOS Press, 2020.

[8] Robert Crystal-Ornelas, Charuleka Varadharajan, Ben Bond-Lamberty, Kristin Boye, Madison Burrus, Shreyas Cholia, Michael Crow, Joan Damerow, Ranjeet Devarakonda, Kim S Ely, et al. A guide to using github for developing and versioning data standards and reporting formats. *Earth and Space Science*, 8(8), 2021.

[9] Joel Cummings and Deborah Stacey. Lean ontology development : An ontology development paradigm based on continuous innovation. In *Knowledge Engineering and Ontology Development*, pages 365–372, 2018.

[10] Antonio De Nicola and Michele Missikoff. A lightweight methodology for rapid ontology engineering. *Communications of the ACM*, 59(3) :79–86, 2016.

[11] ETSI. SmartM2M ; Guidelines for consolidating SAREF with new reference ontology patterns, based on the experience from the ITEA SEAS project. ETSI Technical Report 103 549 V1.1.1., 07 2019.

[12] ETSI. SmartM2M ; Extension to SAREF ; Part 1 : Energy Domain. ETSI Technical Specification 103 410-1 V1.1.2., 05 2020.

[13] ETSI. SmartM2M ; Extension to SAREF ; Part 10 : Water Domain. ETSI Technical Specification 103 410-10 V1.1.1., 07 2020.

[14] ETSI. SmartM2M ; Extension to SAREF ; Part 2 : Environment Domain. ETSI Technical Specification 103 410-2 V1.1.2., 05 2020.

[15] ETSI. SmartM2M ; Extension to SAREF ; Part 4 : Smart Cities Domain. ETSI Technical Specification 103 410-4 V1.1.2., 05 2020.

[16] ETSI. SmartM2M ; Extension to SAREF ; Part 5 : Industry and Manufacturing Domain. ETSI Technical Specification 103 410-5 V1.1.2., 05 2020.

[17] ETSI. SmartM2M ; Extension to SAREF ; Part 6 : Smart Agriculture and Food Chain Domains. ETSI Technical Specification 103 410-6 V1.1.2., 05 2020.

[18] ETSI. SmartM2M ; Extension to SAREF ; Part 7 : Automotive Domain. ETSI Technical Specification 103 410-7 V1.1.1., 07 2020.

[19] ETSI. SmartM2M ; Extension to SAREF ; Part 8 : eHealth/Ageing-well Domain. ETSI Technical Specification 103 410-8 V1.1.1., 07 2020.

[20] ETSI. SmartM2M ; Extension to SAREF ; Part 9 : Wearables Domain. ETSI Technical Specification 103 410-9 V1.1.1., 07 2020.

[21] ETSI. SmartM2M ; SAREF consolidation with new reference ontology patterns, based on the experience from the SEAS project. ETSI Technical Specification 103 548 V1.1.2., 06 2020.

[22] ETSI. SmartM2M ; SAREF Development Framework and Workflow, Streamlining the Development of SAREF and its Extensions. ETSI Technical Specification 103 673 V1.1.1., 2020.

[23] ETSI. SmartM2M ; Smart Applications ; Reference Ontology and oneM2M Mapping. ETSI Technical Specification 103 264 V3.1.1., 02 2020.

[24] ETSI. SmartM2M ; Extension to SAREF ; Part 11 : Lift Domain. ETSI Technical Specification 103 410-11 V1.1.1., 07 2021.

[25] Alba Fernández-Izquierdo and Raúl García-Castro. Themis : a tool for validating ontologies through requirements. In *Software Engineering and Knowledge Engineering*, pages 573–753, 2019.

[26] Mariano Fernández-López, Asunción Gómez-Pérez, and Natalia Juristo. Methontology : from ontological art towards ontological engineering. 1997.

[27] Aldo Gangemi. Ontology design patterns for semantic web content. In *International semantic web conference*, pages 262–276. Springer, 2005.

[28] Aldo Gangemi and Valentina Presutti. Ontology design patterns. In *Handbook on ontologies*, pages 221–243. Springer, 2009.

[29] Asunción Gómez-Pérez and Mari Carmen Suárez-Figueroa. Neon methodology : scenarios for building networks of ontologies. In *16th International Conference on Knowledge Engineering and Knowledge Management Knowledge Patterns (EKAW 2008). Conference Poster*, 2008.

[30] Rafael S Gonçalves, Bijan Parsia, and Ulrike Sattler. Ecco : A hybrid diff tool for owl 2 ontologies. In *OWLED*, volume 849. Citeseer, 2012.

[31] Bernardo Cuenca Grau, Bijan Parsia, Evren Sirin, and Aditya Kalyanpur. Modularity and web ontologies. In *Knowledge Representation*, pages 198–209, 2006.

[32] Michael Grüninger and Mark S Fox. Methodology for the design and evaluation of ontologies. 1995.

[33] Lavdim Halilaj, Irlán Grangel-González, Maria-Esther Vidal, Steffen Lohmann, and Sören Auer. Proactive prevention of false-positive conflicts in distributed ontology development. In *Knowledge Engineering and Ontology Development*, pages 43–51, 2016.

[34] Lavdim Halilaj, Niklas Petersen, Irlán Grangel-González, Christoph Lange, Sören Auer, Gökhan Coskun, and Steffen Lohmann. Vocol : An integrated environment to support version-controlled vocabulary development. In *European Knowledge Acquisition Workshop*, pages 303–319. Springer, 2016.

[35] Maia Hristozova and Leon Sterling. An extreme method for developing lightweight ontologies. In *In Workshop on Ontologies in Agent Systems, 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems*. Citeseer, 2002.

[36] IEEE. *Guide to the Software engineering body of knowledge v3.0*. IEEE Computer society, 2014.

[37] ISO/IEC/IEEE. Systems and software engineering — vocabulary. Standard ISO/IEC/IEEE 24765:2017, 2017.

[38] Alba Fernández Izquierdo. Ontology verification based on lexico-syntactic patterns, November 2020.

[39] Rebecca C Jackson, James P Balhoff, Eric Douglass, Nomi L Harris, Christopher J Mungall, and James A Overton. Robot : a tool for automating ontology workflows. *BMC bioinformatics*, 20(1) :1–10, 2019.

[40] Haruhiko Kaiya and Motoshi Saeki. Using domain ontology as domain knowledge for requirements elicitation. In *14th IEEE International Requirements Engineering Conference (RE'06)*, pages 189–198. IEEE, 2006.

[41] Holger Knublauch and Dimitris Kontokostas. Shapes Constraint Language (SHACL). W3C Recommendation, W3C, July 20 2017.

[42] Petr Kremen, Marek Smid, and Zdenek Kouba. Owl-diff : A practical tool for comparison and merge of owl ontologies. In *2011 22nd International Workshop on Database and Expert Systems Applications*, pages 229–233. IEEE, 2011.

[43] Bernd Krieg-Brückner and Till Mossakowski. Generic ontologies and generic ontology design patterns. In *WOP@ ISWC*, 2017.

[44] Bernd Krieg-Brückner, Till Mossakowski, and Mihai Codescu. Generic ontology design patterns : Roles and change over time. *Advances in Pattern-Based Ontology Engineering*, 51 :25, 2021.

[45] Maxime Lefrançois. Planned ETSI SAREF extensions based on the W3C&OGC SOSA/SSN-compatible SEAS ontology paaerns. In *Workshop on semantic interoperability and standardization in the IoT, SIS-IoT*, page 11p, 2017.

[46] Maxime Lefrançois, Antoine Zimmermann, and Noorani Bakerally. A SPARQL extension for generating RDF from heterogeneous formats. In *European Semantic Web Conference*, pages 35–50. Springer, 2017.

[47] Nicolas Matentzoglu, Chris Mungall, and Damien Goutte-Gattat. Ontology development kit, July 2021. If you use this software, please cite it as below.

[48] Boris Motik, Peter F Patel-Schneider, Bijan Parsia, Conrad Bock, Achille Fokoue, Peter Haase, Rinke Hoekstra, Ian Horrocks, Alan Ruttenberg, Uli Sattler, et al. Owl 2 web ontology language : Structural specification and functional-style syntax. W3c recommendation, W3C, 2009.

[49] Mark A Musen. The protégé project : a look back and a look forward. *AI matters*, 1(4) :4–12, 2015.

[50] Natalya F Noy, Deborah L McGuinness, et al. Ontology development 101 : A guide to creating your first ontology, 2001.

[51] Natalya Fridman Noy, Mark A Musen, et al. Prompt-diff : A fixed-point algorithm for comparing ontology versions. *AAAI/IAAI*, 2002 :744–750, 2002.

[52] David Osumi-Sutherland, Melanie Courtot, James P Balhoff, and Christopher Mungall. Dead simple owl design patterns. *Journal of biomedical semantics*, 8(1) :1–7, 2017.

[53] Silvio Peroni. Samod : an agile methodology for the development of ontologies. In *Proceedings of the 13th OWL : Experiences and Directions Workshop and 5th OWL reasoner evaluation workshop (OWLED-ORE 2016)*, pages 1–14, 2016.

[54] Helena Sofia Pinto, Steffen Staab, and Christoph Tempich. Diligent : Towards a fine-grained methodology for distributed, loosely-controlled and evolving engineering of ontologies. In *ECAI*, volume 16, page 393. Citeseer, 2004.

[55] María Poveda-Villalón, Alba Fernández-Izquierdo, Mariano Fernández-López, and Raúl García-Castro. Lot : An industrial oriented ontology engineering framework. *Engineering Applications of Artificial Intelligence*, 111 :104755, 2022.

[56] María Poveda-Villalón, Asunción Gómez-Pérez, and Mari Carmen Suárez-Figueroa. Oops !(ontology pitfall scanner !) : An on-line tool for ontology evaluation. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 10(2) :7–34, 2014.

[57] Omar Qawasmeh, Maxime Lefrançois, Antoine Zimmermann, and Pierre Maret. Pitfalls in networked and versioned ontologies. In *Knowledge Discovery, Knowledge Engineering and Knowledge Management - 11th International Joint Conference, IC3K 2019, Vienna, Austria, September 17-19, 2019, Revised Selected Papers*, volume 1297 of *Communications in Computer and Information Science*, pages 185–212. Springer, 2019.

[58] Lila Rao, Han Reichgelt, and Kweku-Muata Osei-Bryson. Knowledge elicitation techniques for deriving competency questions for ontologies. In *International Conference on Enterprise Information System*, pages 105–110, 2008.

[59] Amir Azim Sharifloo and Mehrnoush Shamsfard. Using agility in ontology construction. In *Formal Ontologies Meet Industry*, volume 174 of *Frontiers in Artificial Intelligence and Applications*, pages 109–119. IOS Press, 2008.

[60] Martin G Skjæveland, Henrik Forssell, Johan W Klüwer, Daniel Lupp, Evgenij Thorstensen, and Arild Waaler. Pattern-based ontology design and instantiation with reasonable ontology templates. *A Higher-Level View of Ontological Modeling*, page 69, 2019.

[61] Steffen Staab, Rudi Studer, H-P Schnurr, and York Sure. Knowledge processes and ontologies. *IEEE Intelligent systems*, 16(1) :26–34, 2001.

[62] Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, and Boris Villazón-Terrazas. How to write and use the ontology requirements specification document. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 966–982. Springer, 2009.

[63] Akkharawoot Takhom, Sasiporn Usanavasin, Thepchai Supnithi, and Prachya Boonkwan. A collaborative framework supporting ontology development based on agile and scrum model. *IEICE TRANSACTIONS on Information and Systems*, 103(12) :2568–2577, 2020.

[64] Mike Uschold and Michael Gruninger. Ontologies : Principles, methods and applications. *The knowledge engineering review*, 11(2) :93–136, 1996.

[65] Max Volkel, Wolf Winkler, York Sure, S Ryszard Kruk, and Marcin Synak. Semversion : A versioning system for rdf and ontologies. In *Proc. of ESWC*, 2005.

eCISE-OWL : Représentation OWL du Schéma de Données de l'Environnement Commun d'Échange d'Information pour le Domaine Maritime

Nathalie Aussenac-Gilles, Catherine Comparot, Antoine Dupuy, Nabil El Malki, Ronan Tournier, Ba-Huy Tran, Cassia Trojahn

[1] IRIT, Université de Toulouse, CNRS, UT1, UT2, Toulouse, France

prenom.nom ou prenom.nom-composé@irit.fr

Résumé

L'Environnement Commun de Partage d'Information (CISE) est un modèle de donnée qui résulte d'une initiative collaborative pour promouvoir le partage automatisé d'informations entre les autorités de surveillance maritime. L'exploitation de CISE est cependant limitée par sa sérialisation via uniquement un schéma XML, peu adapté pour fournir une sémantique plus riche, une interopérabilité sémantique et des capacités de raisonnement. La transformation de la version étendue de CISE (eCISE) en ontologie est indispensable pour faciliter l'accès aux données d'échanges maritimes selon les principes de l'Ontology Based Data Access, et le raisonnement sur ces données. Cet article présente eCISE-OWL, une représentation OWL de eCISE, désormais disponible en accès ouvert. Pour obtenir cette ontologie, nous proposons un processus de transformation de schémas XML (en XSD) vers OWL qui améliore l'état de l'art, dont le code est rendu public et qui comporte une étape de validation par des experts du domaine.

Mots-clés

CISE, ontologie, domaine maritime, interopérabilité

Abstract

The Common Information Sharing Environment (CISE) is the result of a collaborative initiative to promote automated information sharing among maritime surveillance authorities. L'exploitation de CISE is however limited by its serialization of CISE via only an XML schema, insufficient to provide richer semantics, semantic interoperability and reasoning capabilities. This paper presents eCISE-OWL, an OWL representation of the extended version of CISE (eCISE). eCISE-OWL is the result of an improved process of transforming XML schemas (XSD) into OWL and of validation and correction efforts by domain experts. We discuss the use of eCISE-OWL in maritime exchange data access (OBDA) and reasoning tasks.

Keywords

CISE, ontology, maritime domain, interoperability

1 Introduction

Rendre les systèmes de surveillance maritime interopérables est crucial pour la coopération entre les pays, en particulier en cas de crises maritimes dans des zones frontalières entre pays. Dans cet objectif, l'hétérogénéité entre les système nationaux et les structures de données des différents acteurs soulèvent de nombreux problèmes. Afin de permettre aux autorités maritimes d'échanger des informations de manière automatique et sécurisée, elles ont adopté un environnement commun de partage d'informations (CISE) [1]. Il fournit un cadre décentralisé et un modèle de données pour l'échange d'informations point à point entre les secteurs et les frontières. Il implique plus de 300 autorités européennes et nationales ayant des responsabilités en matière de surveillance maritime, et effectuant de nombreuses tâches de surveillance opérationnelle. Ces autorités bénéficient directement d'être connectées au réseau CISE, pour des objectifs aussi divers que la sûreté et la sécurité du transport maritime, le contrôle de la pêche, la pollution, et la défense. Depuis 2014, CISE est retenu pour soutenir la mise en œuvre de la stratégie de sécurité maritime de l'Union Européenne (EUMSS).

L'adoption du modèle de données CISE [2] et de ses différentes versions – en particulier, *Extended-CISE* (eCISE) [1] – est cependant limitée par sa sérialisation via un schéma XML uniquement, dont la sémantique n'est pas assez riche pour garantir une interopérabilité sémantique des données ou pour servir de support à un raisonnement. Or ces fonctionnalités s'avèrent très utiles pour associer des données venant de différentes sources de données CISE, les vérifier ou encore en déduire de nouveaux éléments. Un premier effort dans cette direction a été proposé par le projet européen ROBORDER [14], qui a généré une représentation ontologique du modèle de données CISE tel qu'il a été défini dans le projet EUCISE2020 [5], en convertissant le modèle UML en OWL. Même si l'ontologie et son processus de construction sont décrits (incomplètement) dans l'article cité, aucun

1. http://www.emsa.europa.eu/cise.html
2. http://emsa.europa.eu/cise-documentation/cise-data-model-1.5.3/

des deux n'est disponible publiquement.
Bien que le passage de données XML ou de schémas XSD à une représentation sémantique (RDF, RDFS ou OWL) soit une question étudiée de longue date dans le domaine du Web sémantique, une transformation automatique simple est rarement correcte. Ce processus se heurte à la difficulté de gérer des noeuds anonymes, de traiter la représentation de noeuds complexes, de capturer la sémantique des balises purement structurelles, ou encore de produire des constructeurs liés à la structuration [12, 10, 4, 22].
Ce travail est réalisé dans le cadre du projet H2020 EFFECTOR [3], qui veut proposer un cadre d'interopérabilité et des services de fusion et d'analyse de données associés pour la surveillance maritime et la sécurité des frontières. Ainsi, EFFECTOR vise à améliorer le processus d'aide à la décision et à favoriser la collaboration des acteurs maritimes au niveau local, régional et transnational. Le modèle de données CISE joue un rôle essentiel dans le projet car les messages échangés par les différents acteurs sont basés sur les diverses versions de ce modèle. Cependant, ces messages sont difficilement accessibles car stockés dans plusieurs bases de données internes. Un système d'accès aux données via une ontologie (OBDA) va donc être mis en place afin de contribuer à l'interoperabilité des systèmes et de faciliter les échanges de données entre partenaires. De plus, afin d'aider l'opérateur chargé de la surveillance maritime, l'ontologie permettra de produire des inférences et de générer de nouveaux faits signalant de potentielles anomalies.
Cet article présente eCISE-OWL, une représentation OWL du modèle eCISE. eCISE étend le modèle CISE en améliorant le vocabulaire maritime de CISE et en élargissant sa portée à la surveillance terrestre et à l'échange d'informations opérationnelles. Outre l'ontologie elle-même, notre contribution réside aussi dans le processus original de transformation d'un schéma XML (XSD) en OWL. Ce processus tient compte des particularités du modèle de données CISE pour générer un premier modèle OWL, que des experts du domaine ensuite valident et corrigent. Il intègre et étend des travaux existants, décrits de manière formelle dans plusieurs articles mais qui n'avaient jamais été regroupés dans une unique chaîne de traitement et qui n'étaient pas réutilisables. Nous avons amélioré significativement la gestion des collections, représentées jusque là comme des collections de `owl:oneOf`, ce qui ralentit les performances au moment de requêter ces données. De plus, nous avons simplifié la représentation des classes d'association, qui sont très nombreuses dans eCISE. Enfin, aussi bien le code de génération de l'ontologie que l'ontologie [4] elle-même sont disponibles publiquement.
Nous poursuivons cet article en présentant un état de l'art sur les ontologies maritimes en lien avec le modèle de données CISE et ses variantes, ainsi que sur les solutions existantes pour générer des modèles OWL à partir de schémas XSD (section 2. La section 3 introduit le modèle CISE et son extension eCISE. Les cas d'usages de l'ontologie sont ensuite présentés (Section 4). Nous détaillons la méthodologie de construction en Section 5. L'ontologie ainsi que son évaluation sont présentés dans la Section 6. Finalement, la Section 7 conclut l'article et dessine les perspectives pour les travaux futurs.

2 Travaux liées

Ontologies maritimes La surveillance maritime automatisée reçoit un intérêt particulier depuis quelques années, ce qui est attesté par un grand nombre de projets abordant les différents défis du domaine (ROBORDER, EUCISE2020, ANDROMEDA, MARISA, datAcron, et CoopP, pour en citer quelques uns). Les technologies sémantiques ont prouvé leur pertinence en facilitant le partage automatisé d'informations entre les systèmes de surveillance maritime. En particulier, dans le cadre du projet datAcron, une ontologie a été proposée pour représenter des trajectoires d'objets en mouvement[15]. Plus proche de nos travaux, dans le cadre du projet ROBORDER, l'ontologie EUCISE-OWL résulte [14] d'une conversion UML en OWL du modèle de données EUCISE2020 (une extension du modèle CISE Data Model v1.0 [5] pour couvrir des sources de données supplémentaires). EUCISE-OWL a été la première tentative d'exploiter entièrement le schéma CISE pour développer une ontologie qui facilite l'échange d'informations dans le domaine maritime. Cependant, la ressource et son processus de construction ne sont pas publiquement disponibles.
D'autres ontologies maritimes se trouvent dans la littérature. Les travaux de [8] proposent une représentation ontologique des différents types de navires et des paramètres pertinents, en fonction de l'AIS (Automatic Identification System), dans le cadre de la tâche d'analyse de trafic maritime. Dans [6], l'analyse de trajectoires sémantiques et les localisations géographiques des objets maritimes est réalisée grâce à des ontologies de domaine. L'approche combine le traitement de données statiques et en temps réel provenant de différentes sources à l'aide de techniques d'accès aux données basées sur l'ontologie (OBDA). Un autre aspect du domaine concernant la détection ou la prédiction de comportements anormaux des navires, plusieurs travaux ont utilisé des ontologies dédiées à cette tâche. Dans [19, 20], les ontologies spatiales et la représentation sémantique des trajectoires servent à caractériser le comportement anormal des navires, sur la base de propriétés sémantiques formelles servant à raisonner sur les données. Alors que ces méthodes reposent principalement sur des ontologies créées manuellement ou dérivées de ressources non ontologiques comme des schémas XML ou des diagrammes UML, d'autres ontologies maritimes ont été construites à partir de textes [24]. Finalement, sous un autre angle, une ontologie de la réglementation maritime a été définie décrivant par exemple les procédures portuaires et la maintenance de navires, afin d'évaluer l'impact des nouveaux règlements et de retracer leur origine législative [11].

Passage XML/XSD vers OWL La représentation sémantique (RDF ou OWL) de données XML ou de schémas XSD

3. https://www.effector-project.eu/
4. L'ontologie est disponible https://www.irit.fr/recherches/MELODI/ontologies/ecise/index-en.html

est une question étudiée de longue date dans le domaine du Web sémantique. Cependant, les balises ne se situant pas toutes au même niveau d'abstraction, une transformation automatique simple est rarement efficace et correcte. Lorsque les éléments définis entre balises sont eux mêmes complexes, ils relèvent de plusieurs types en lien représentés par de multiples propriétés. Ils peuvent même contribuer à la fois à enrichir l'ontologie et à la peupler. C'est le cas de l'exploitation de fiches XML dans le projet MOANO par exemple [3]. Dans tous les cas, le passage du XSD à des classes OWL ou des types RDF se heurte à la difficulté de gérer des noeuds anonymes, de traiter la représentation de noeuds complexes, des énumérations, de gérer les types XSD, ou encore les balises liées à la structuration et non sémantiques.

Plusieurs approches sont mises en oeuvre dans les logiciels de l'état de l'art. Un premier ensemble d'outils, dits de "lifting", convertissent un schéma XML (XSD) en un schéma RDF, tel que RDFS ou OWL. C'est le cas de XML2OWL (qui part d'un document XML ou d'un XSD) ou de XSD2OWL [5] (à partir d'un XSD). L'approche la plus classique consiste à utiliser XSLT, le langage de transformation de schémas XSD, en considérant RDF-XML comme un schéma XML particulier [6]. XSLT est d'ailleurs à la base de GRDDL [7], un mécanisme qui permet de rajouter des balises dans un document XML pour indiquer que les données décrites peuvent être traduites vers RDF à l'aide de XSLT. Cependant, un wiki du W3C [8] souligne que si XSLT est adapté pour convertir une majorité de modèles XML en RDF, il se heurte à plusieurs limites : il génère des modèles pas optimisés et illisibles par un humain ; dans le cas de modèles complexes, XSLT ne sait pas traiter toutes les situations, par exemple des structures imbriquées ou des textes longs entre balises. C'est en effet le cas de l'outil Ontmalizer [9] [23] qui fournit l'ensemble de règles de transformations adoptées dans cet article, mais qui génère un résultat de conversion complexe. Le site MIT Simile fournissait une liste de quelques autres "RDF-izers" [10] dont la plupart ne sont plus accessibles ou ne gèrent pas le format XML. Le W3C dresse une autre liste [11] d'outils de conversion de XML vers RDF, dont TopBraid Composer (un logiciel commercial) avec un plugin qui gère le passage de XSD vers OWL ; KREXTOR, une plateforme qui sait traiter plusieurs variantes de XML à l'aide de transformations XSLT ; Rhizomik, qui fait appel à XML2RDF et XSD2RDF ; GRDDL ; XHTML, etc.

Une alternative performante à XSLT repose sur des règles RML, dont le pouvoir d'expression est plus puissant à traiter des schémas complexes, et dont le format se prête bien à une reformulation lisible par un humain. Le logiciel SDM-RDFIZER [12] s'appuie sur RML pour traiter une grande diversité de formats. Son avantage est le passage à l'échelle, les règles étant optimisées pour traiter de grands volumes de données.

Malheureusement, les outils présentés dans les articles de recherche (comme les deux que nous venons de citer) sont rarement accessibles.

3 Schémas CISE et eCISE

3.1 De CISE à eCISE

Le modèle de données CISE a pour ambition de servir de format pour le partage d'information de surveillance maritime entre secteurs et pays. Dans cette optique, il décrit sept entités principales (`Agent`, `Object`, `Location`, `Document`, `Event`, `Risk`, `Period`) et onze entités secondaires (`Vessel`, `Cargo`, `Operational Asset`, `Person`, `Organization`, `Movement`, `Incident`, `Anomaly`, `Action`, `Unique Identifier`, `Metadata`). Ce modèle permet aux différentes autorités de bénéficier d'un vocabulaire commun pour décrire notamment les événements observés. Il est décliné dans les formalismes RDF et XSD. Sur la Figure 1, les entités du modèle CISE correspondent aux hexagones non colorés.

Le modèle de données eCISE enrichit le vocabulaire de CISE pour les domaines maritime et terrestre.Il fournit un ensemble plus riche de types de navires, d'informations fournies par le système d'identification de navires AIS (Automatic Identification System) et de capteurs radar ; il liste également un ensemble plus complet d'anomalies et de règles maritimes, mais aussi terrestres, avec un nombre important de types pour chacune de ses entités. Ce modèle est construit sur la dernière version du modèle de données CISE utilisé dans le projet EUCISE2020 [9]. Sur la Figure 1, les entités centrales de eCISE qui complètent celles du modèle CISE, correspondent aux hexagones de couleur.

3.2 Le modèle XML d'eCISE

Les modèles de données CISE et eCISE sont décrits dans un document de spécification (diagrammes de classes UML) et implémentés en XSD (schémas XML). Les fichiers XSD de CISE ont été produits à partir de transformations, i.e. un ensemble de règles de correspondance indiquant comment générer des éléments XSD à partir des éléments UML. Les choix effectués lors de ce processus impactent la structure XSD obtenue et doivent être pris en compte lors de la génération de sa représentation OWL. En ce qui concerne CISE, chaque fichier .xsd représente une ou plusieurs entités du modèle, où chaque entité est représentée via une balise `xs:complexType`. La notion de hiérarchie de spécialisation entre entités est représentée par la balise `xs:extension`. Les éléments (au sens XML) qui composent plus spécifiquement une entité sont alors décrits en XSD au sein de la balise

5. https://gist.github.com/pebbie/5704765
6. https://rml.io/docs/rml/tutorials/xml/
7. Gleaning Resource Descriptions from Dialects of Languages https://www.w3.org/TR/grddl/
8. https://www.w3.org/community/rax/wiki/XML_to_RDF_Transformation_processes_using_XSLT
9. https://www.w3.org/wiki/HCLSIG/Tools\#Ontmalizer
10. https://en.wikipedia.org/wiki/List_of_SIMILE_projects#RDFizer
11. https://www.w3.org/wiki/ConverterToRdf#XML
12. https://pypi.org/project/rdfizer/

FIGURE 1 – Aperçu des vocabulaires du modèle de base CISE (entités en blanc) et du modèle eCISE (entités CISE enrichies par les entités en couleur).

`xs:complexContent`. Ils sont listés dans l'ordre où ils doivent apparaître dans une balise `xs:sequence`. Chacun de ces éléments correspond soit à une propriété propre à l'entité, soit à une association avec une autre entité. L'extrait ci-dessous décrit l'entité `Vehicle` comme une sous-classe de `Object` liée à l'entité Cargo par une association représentée en XML par la propriété `CargoRel`; la valeur explicite 0 de `xs:minOccurs` et la valeur implicite 1 de `xs:maxOccurs` indiquent une cardinalité de 0,1 ; la propriété est donc facultative et ne peut apparaître qu'une fois. Il est à noter que `Cargo` désigne une cargaison (i.e., un ensemble de marchandises transporté par un véhicule entre deux ports), et non un type de bateau.

```
<xs:complexType name="Vehicle" abstract="true">
 <xs:annotation>
  <xs:documentation>
   The Vehicle is a sub-class of Object [...]
  </xs:documentation>
 </xs:annotation>
 <xs:complexContent>
  <xs:extension base="object:Object">
   <xs:sequence>
    <xs:element name="CargoRel" minOccurs="0">
     <xs:complexType>
      <xs:complexContent>
       <xs:extension base="rel:Relationship">
        <xs:sequence>
         <xs:element name="Cargo" type="cargo:Cargo"
                     minOccurs="0"/>
        </xs:sequence>
       </xs:extension>
      </xs:complexContent>
     </xs:complexType>
    </xs:element>
   </xs:sequence>
  </xs:extension>
 </xs:complexContent>
</xs:complexType>
```

4 Exigences en matière d'ontologie

Chaque système informatique employé dans les centres de coordination et de sauvetage maritime ayant son propre vocabulaire et son propre schémas de données, une ontologie commune offre un modèle pivot de haut niveau qui, d'un côté, facilite l'interopérabilité et, de l'autre, permet d'envisager un raisonnement sur les données de surveillance. Le modèle CISE représentant le vocabulaire de la surveillance maritime, dans le cadre d'EFFECTOR, nous avons retenu ce modèle comme base pour construire cette ontologie de domaine. Nous décrivons dans ce qui suit les deux services retenus dans EFFECTOR qui justifient l'utilisation d'une ontologie pivot.

4.1 Ontology Based Data Access

L'*Ontology Based Data Access* (OBDA) [7] est un paradigme d'accès aux données à travers une couche conceptuelle. Habituellement, cette couche est spécifiée via un schéma RDF ou une ontologie OWL et les données sont stockées dans des bases relationnelles. Les termes du niveau conceptuel sont associés à la couche de données via un mapping associant individuellement chaque élément de la couche conceptuelle à une requête SQL (potentiellement complexe) sur les sources de données relationnelles. Le graphe virtuel ainsi obtenu peut être requêté via un langage de requête sur des données RDF, tel que SPARQL.

Les acteurs de la surveillance et du sauvetage maritime reçoivent les message CISE via le réseau de noeuds CISE des états membres. Les informations échangées sont stockées dans les bases de données locales à chaque système national, qui gèrent également d'autres données en provenance d'autres sources (AIS, Sat-AIS, radar ...). Comme le projet EFFECTOR vise à améliorer le processus de prise de décision et à favoriser la collaboration entre les acteurs, échanger des données les plates-formes est un enjeu clé. Dans cet objectif, une solution adaptée repose sur l'approche OBDA où le niveau conceptuel fait appel à une ontologie construite sur un vocabulaire commun trans-frontalier et multi-plateformes.

4.2 Raisonnement sur des messages CISE

Une autre utilisation de l'ontologie concerne la capacité de raisonnement afin d'inférer de nouveaux éléments à partir de l'instanciation de parties spécifiques de l'ontologie, selon les différents scénarios définis dans le projet. Il s'agit notamment d'offrir un support à l'opérateur humain dans le processus de détection des anomalies. Par souci de confidentialité, nous ne pouvons communiquer les détails des scénarios et les exemples suivants sont des cas classiques du domaine [18, 20].

Alerte vitesse Si un bateau se déplace à une vitesse supérieure à la vitesse autorisée pour le type de bateau en question, une alerte vitesse doit être générée. En s'inspirant de [20], cette alerte peut être produite via une règle SWRL adaptée au vocabulaire eCISE, comme suit :

```
ObjectLocation(?objectLocation),
hasInvolvedLocation(?objectLocation,
```

```
?location), hasInvolvedObject(?objectLocation,
?vehicle), Vessel(?vehicle),
speed(?objectLocation,?speed),
maximumspeed(?vehicule,?maxSpeed),
greaterThan(?speed,?maxSpeed)
→ MaritimeAnomaly(?anomaly),
hasMaritimeAnomalyType(?anomaly,:HighSpeed)
```

Risque de pollution S'il existe une risque de collision imminente entre deux bateaux et qu'au moins l'un des navires impliqués a une cargaison dangereuse, alors il y a un risque de pollution :

```
ObjectEvent(?objectEvent),
hasInvolvedEvent(?objectEvent,
:VesselImminentCollision),
hasInvolvedObject(?objectEvent,?vehicle),
Vessel(?vehicle), hasCargo(?vehicle,?cargo),
hasContainedCargoUnit(?cargo,?containmentUnit),
hasPollutionCode(?containmentUnit,?pollution-
CodeType) → Risk(?risk),
hasRiskType(?risk,:Pollution)
```

5 Réingénierie de ressources non ontologiques

La méthodologie de construction de l'ontologie eCISE-OWL est conforme au Scénario 2 *Réutilisation et réingénierie des ressources non ontologiques (RNO)* de la méthodologie NeOn [16, 21]. Le processus de réingénierie des RNO a été défini pour transformer les RNO en ontologies. Dans notre cas, les RNO correspondent aux schémas XSD (décrits en XML) : ils sont homogènes dans leur modèle de données. En [21], les auteurs suggèrent de se placer au niveau conceptuel pour étudier les correspondances entre le modèle source et sa conversion RDF, en explicitant comment chaque élément du schéma XML doit être traduit en RDF ou OWL. La mise en forme (manuelle) de ces règles pour former un patron de conversion est une étapes clé de la traduction, qui permet d'automatiser ensuite la traduction de données décrites selon ce modèle. Dans notre cas, le problème est de transformer un schéma, et de traiter de la même manière chaque structure similaire dans ce schéma, selon les mêmes règles.

5.1 De XML à OWL

Faute de disposer de logiciel opérationnel et libre pour transformer les modèles XML en RDF qui sachent traiter la représentation des relations n-aires entre classes, nous avons implémenté des règles de conversion de XML vers RDF qui répondent aux exigences du modèle CISE. Il nous a paru souhaitable que ce processus produise aussi automatiquement des descriptions des éléments (via `rdfs:comment` par exemple) à partir de la documentation. Afin de traiter le modèle CISE, les difficultés à dépasser sont notamment liées au nombre important d'énumérations (types possibles d'entités spécifiques) à convertir, au manque d'information quant au nommage des classes d'association et à la conversion des contraintes de cardinalité.

Pour répondre à ces besoins, nous avons développé un processus de transformation. Il s'inspire du travail de [2] et utilise un ensemble de règles de transformation. Ce processus, implémenté en Python, utilise `rdflib`. Il parcourt des sources XSD et des sources externes afin d'extraire les éléments nécessaires à la construction de l'ontologie. Pour chaque type d'élément XSD, une règle de transformation correspondante est appliquée vers le formalisme OWL. Le script récupère la documentation du modèle de données eCISE pour en extraire les commentaires qui serviront à documenter les entités de l'ontologie (`rdfs:label` et `rdfs:comment`). Pour cette étape d'enrichissement terminologique de l'ontologie, nous utilisons comme source externe le livrable D3.1 d'Andromeda [1] au format PDF. Notre outil utilise la librairie PDFReader et effectue une lecture page à page de ce PDF. Cette étape est à généraliser afin de rentre le processus capable de traiter d'autres sources de données.

5.2 Règles de transformation

Les règles de transformations adoptées ici reprennent la plupart des règles de transformation de l'outil Ontmalizer [13] [23]. Concernant les classes d'association UML et les énumérations, nous avons adapté ces règles en suivant la proposition de [14]. La table 1 liste les correspondances entre les éléments XSD et les définitions OWL ainsi que des éléments définis pour le traitement de certains types spécifiques tels que les énumérations et les classes d'association. Des exemples de transformation sont introduits par la suite, pour les principaux type de constructeurs OWL.

Espaces de noms Les fichiers sources XSD indiquent des espaces de nom par groupe d'entités. Ces espaces de nom ont été retranscrits tels quels dans l'ontologie pour respecter les conventions du modèle de données eCISE. La version actuelle de l'ontologie eCISE-OWL comporte ainsi des espaces de nom tels que `event`, `vessel`, `object` et `location`.

Classes Toute entité ou classe décrite dans le modèle de données eCISE est une sous-classe de `:Entity` (sous-classe de `:owl:Thing`). A titre d'exemple, nous présentons ci-dessous les représentations XML et OWL de l'entité `:Vessel` :

```
<xs:complexType name="Vessel">
 <xs:complexContent>
  <xs:extension base="object:Vehicle">
   <xs:sequence>
    [...]
   </xs:sequence>
  </xs:extension>
 </xs:complexContent>
</xs:complexType>

[:Vessel
    rdf:type owl:Class ;
    rdfs:subClassOf :Vehicle ;
    rdfs:comment "The class Vessel is a
     sub-class of the class Vehicle. A vessel refers
     to a ship or a boat. [...]" ;
    rdfs:label "Vessel" .
]
```

13. https://www.w3.org/wiki/HCLSIG/Tools\#Ontmalizer

Élément XSD	*Définitions OWL*
xs :simpleType	rdfs :Datatype
xs :simpleType	rdfs :Datatype
xs :enumeration	owl :Class et owl :Individual
xs :complexType over xs :complexContent	owl :Class
xs :complexType over xs :simpleContent	owl :Class
xs :element (global) with complex type	owl :Class and rdfs :subclassOf
xs :element (global) with simple type	owl :Datatype
xs :element (local to a type)	owl :DatatypeProperty or owl :ObjectProperty
xs :group	owl :Class
xs :attributeGroup	owl :Class

TABLE 1 – Règles de conversion XSD en OWL.

Propriétés Les classes sont liées à d'autres types de données (soit des classes, soit des valeurs). Pour ce type d'élément, nous avons suivi les bonnes pratiques de nommage introduites en [17] (à noter les noms de propriétés de classes préfixés par un 'has'). Dans l'exemple, la propriété de classe `:hasCargo` associe `:Vehicle` et `:Cargo`, selon l'extrait du schéma XSD reproduit en Section 3.2.

```
<xs:complexType name="Vehicle" abstract="true">
 <xs:complexContent>
  <xs:extension base="object:Object">
   <xs:sequence>
   [...]
   <xs:element name="CargoRel"  minOccurs="0">
    <xs:complexType>
     <xs:complexContent>
    <xs:extension base="rel:Relationship">
   <xs:sequence>
    <xs:element name="Cargo" type="cargo:Cargo"/>
</xs:sequence>
  </xs:extension>
 </xs:complexContent>
</xs:complexType>
  </xs:element>
 </xs:sequence>
</xs:extension>
</xs:complexContent>
```

```
[:hasCargo
     rdf:type owl:ObjectProperty ;
     rdfs:domain :Vehicle ;
     rdfs:range :Cargo .
]
```

Le renommage de la propriété CargoRel suit la règle de nommage que nous avons définie pour les entités `ObjectProperty` par similarité avec des pratiques en usage dans la communauté [14] : retrait du suffixe Rel et ajout de "has" en début de mot.

Classes d'association Une classe d'association (UML) est un type de classe spécifique qui définit la connexion entre les entités principales du modèle, en utilisant des attributs spécifiques appelés "rôles d'association". Les classes d'association du modèle eCISE héritent systématiquement de la classe `Relationship` que nous avons choisi, inspirés de [14], de renommer `:AssociationClass` dans l'ontologie.

14. https://github.com/G-Node/python-odml/issues/112

```
[
:AssociationClass rdf:type owl:Class ;
  rdfs:comment "Abtract class representing a
    relationship of the CISE data model." ;
  rdfs:label "AssociationClass" .
]
```

Les classes d'association peuvent avoir des propriétés et des types de données supplémentaires qui leur sont propres. Inspirés de [14], nous représentons les classes d'association comme des éléments du type `owl:Class` (et non simplement comme des propriétés d'objet), tandis que les rôles d'association sont représentés par des propriétés de type `owl:ObjectProperty` dont le domaine est la classe d'association.

Dans le fragment XSD ci-dessous, l'entité `Object` possède un élément renvoyant à la classe `Event` et inversement. Les éléments `InvolvedObjectRel` (élément de Event) et `InvolvedEventRel` (élément de Object) possèdent les mêmes attributs, à l'exception de la référence à la classe associée (Object pour Event et inversement). On considère ces deux éléments pour créer une classe d'association regroupant les attributs communs aux deux éléments XSD et les deux références aux classes associées. Dans le cas où il n'y a pas d'attribut en plus des références aux classes associées, la création d'une entité de type `ObjectProperty` est préférée.

```
<xs:complexType name="Object" abstract="true">
 <xs:complexContent>
  <xs:extension base="entity:Entity">
   <xs:sequence>
    <xs:element name="InvolvedEventRel">
     <xs:complexType>
      <xs:complexContent>
       <xs:extension base="rel:Relationship">
        <xs:sequence>
         <xs:element name="Event"
              type="event:Event"/>
         <xs:element name="ObjectRole"
              type="event:ObjectRoleInEventType"/>
        </xs:sequence>
       </xs:extension>
      </xs:complexContent>
     </xs:complexType>
    </xs:element>
    [...]
   </xs:sequence>
  </xs:extension>
```

```
 </xs:complexContent>
</xs:complexType>

<xs:complexType name="Event" abstract="true">
 <xs:complexContent>
  <xs:extension base="entity:Entity">
   <xs:sequence>
    [...]
    <xs:element name="InvolvedObjectRel">
     <xs:complexType>
      <xs:complexContent>
       <xs:extension base="rel:Relationship">
        <xs:sequence>
         <xs:element name="Object"
              type="object:Object"/>
         <xs:element name="ObjectRole"
              type="event:ObjectRoleInEventType"/>
         <xs:element name="InvolvementPeriod"
              type="period:Period"/>
        </xs:sequence>
       </xs:extension>
      </xs:complexContent>
     </xs:complexType>
    </xs:element>
    [...]
   </xs:sequence>
  </xs:extension>
 </xs:complexContent>
</xs:complexType>

[
:ObjectEvent rdf:type owl:Class ;
   rdfs:subClassOf :AssociationClass ;
   rdfs:label "ObjectEvent" .

:hasObject rdf:type owl:ObjectProperty ;
    rdfs:domain event:ObjectEvent ;
    rdfs:range object:Object ;

:hasEvent rdf:type owl:ObjectProperty ;
    rdfs:domain object:ObjectEvent ;
    rdfs:range event:Event ;

:hasInvolvementPeriod rdf:type owl:ObjectProperty ;
    rdfs:domain object:ObjectEvent ;
    rdfs:range period:Period ;
]
```

Enumérations Les énumérations en XML définissent les types possibles d'entités spécifiques. Contrairement aux propositions de conversion des outils existants, qui utilisent le constructeur `owl:oneOf` sur les valeurs possibles, une solution plus élégante consiste à définir une classe `:EnumerationType` dont les valeurs possibles énumérées sont des instances. Inspirés de [14], nous choisissons de représenter les énumérations comme des classes (`owl:Class`) qui possèdent en outre une liste prédéfinie d'instances. La plupart des axiomes de l'ontologie proviennent alors des énumérations du modèle de données.

```
<xs:simpleType name="VesselType">
 <xs:restriction base="xs:string">
  [...]
  <xs:enumeration value="Ferry">
  </xs:enumeration>
  <xs:enumeration value="Fishing">
  </xs:enumeration>
  [...]
 </xs:restriction>
</xs:simpleType

:VesselType rdf:type owl:Class ;
            rdfs:subClassOf :EnumerationType ;
            rdfs:label "VesselType" .

:Ferry rdf:type owl:NamedIndividual, :VesselType ;
       rdfs:label "Ferry" .

:Fishing rdf:type owl:NamedIndividual, :VesselType ;
         rdfs:label "Fishing" .
```

Les énumérations représentent la grande majorité des conversions à traiter. Les requêtes effectuées depuis le point d'accès SPARQL de GraphDB sur l'ontologie eCISE permettent de chiffrer la proportion de classes d'énumération et d'individus les peuplant. Les classes d'association représentent un total de 141 classes sur les 269 de l'ontologie. Ces classes sont peuplées de 16 648 individus.

Contraintes La conversion des contraintes de cardinalité des schémas de données a soulevé des questions sur la pertinence de leur retranscription dans un modèle sémantique géré par l'hypothèse du monde ouvert. En effet, à cause de cette hypothèse, les contraintes de cardinalité ne peuvent indiquer que des maximums ou minimums possibles, mais ne peuvent contraindre une cardinalité multiple. Une autre difficulté vient des valeurs par défaut de ces cardinalités (minOccurs et maxOccurs) en XSD, qui est 1, alors qu'elle doit être explicitée en OWL. Suivant l'utilisation prévue de l'ontologie, les contraintes sont gérées par des systèmes annexes à l'ontologie (à la réception d'un message CISE ou eCISE par exemple) et avant l'utilisation de cette ontologie. Nous proposons d'ajouter une option permettant de choisir si les contraintes de cardinalité doivent être retranscrites ou non lors de la génération des ontologies. Ces contraintes sont représentées par des restrictions OWL sur des éléments de type `owl:ObjectProperty` et `owl:DataProperty`.

```
<xs:complexType name="Vessel">
 <xs:annotation>
  <xs:documentation>The class [...].
  </xs:documentation>
 </xs:annotation>
 <xs:complexContent>
  <xs:extension base="object:Vehicle">
   <xs:sequence>
    [...]
    <xs:element name="ConditionOfTheCargoAndBallast"
    type="vessel:ConditionOfTheCargoAndBallastType" />
    [...]
   </xs:sequence>
  </xs:extension>
 </xs:complexContent>
</xs:complexType>

[ a owl:Restriction ;
  owl:onProperty
vessel:hasConditionOfTheCargoAndBallastMinCardinality ;
  owl:minCardinality 0 .
]
```

5.3 Validation manuelle

La génération de l'ontologie eCISE-OWL à partir du modèle eCISE exige que l'ontologie possède les caractéristiques inscrites dans le modèle de base. Dans cette optique,

il a été nécessaire de vérifier la transformation des propriétés du modèle (entités générées vs. spécifications du modèle du livrable du projet ANDROMEDA [1]). Cette opération a été effectuée manuellement par 3 experts, avec l'aide du logiciel Protege. Pour chaque type d'entité, son étiquette, sa classe et le commentaire associé ont été vérifiés Ce travail de vérification et de correction a nécessité plusieurs ateliers de travail et s'est déroulé sur une vingtaine d'heures.
Au cours de la vérification, les incohérences identifiées impliquent exclusivement les noms des classes d'association. En effet, les étiquettes de ces classes ne sont pas spécifiées dans les fichiers sources .xsd. Deux solutions sont alors possibles pour obtenir un résultat de conversion en accord avec les spécifications : reconstruire le nommage pour qu'il corresponde aux spécifications du modèle de données, en spécifiant dans un fichier le nom de la classe d'association pour deux classes données ; ou alors effectuer la conversion en s'appuyant uniquement sur le nommage spécifié dans les fichiers .xsd. Le premier cas implique une programmation spécifique aux modèles de données CISE et eCISE. Dans le second cas, il est possible de proposer une solution plus générique qui serait utilisable à partir d'autres modèles de données.
Afin de ne pas altérer le modèle de données source l'ontologie doit être conforme à la description du modèle UML des documents PDF), la modification des sources XSD a été écartée des solutions de correction de l'ontologie. Ces corrections ont été apportées par un script Python de renommage ayant pour paramètres les noms de classes incohérents et les noms de classes du livrable ANDROMEDA. Le renommage est effectué de façon itérative.

6 eCISE-OWL

L'ontologie eCISE-OWL contient au total 268 classes, 297 propriétés d'objets et 332 propriétés de données. Les principales métriques de l'ontologie, comparées à celles de l'ontologie EUCISE-OWL, sont résumées dans le Tableau 2. Elles confirment la plus grande exhaustivité de notre ontologie par rapport à la première initiative de création de EUCISE-OWL. La Figure 2 présente la hiérarchie des concepts principaux de l'ontologie.
L'ontologie générée a été évaluée avec différentes métriques. Dans une première évaluation, nous avons utilisé OOPS ! (OntOlogy Pitfall Scanner !)[13] [15], permettant d'évaluer la qualité de modélisation de l'ontologie. Cet outil identifie les erreurs de modélisation selon les dimensions structurelles, fonctionnelles et de profilage de l'utilisabilité. Il fournit également un indicateur (critique, important, mineur) pour chaque écueil identifié, en fonction de l'indice respectif. Dans le cas d'eCISE-OWL, aucun écueil n'a été détecté concernant les dimensions structurelle, fonctionnelle, cohérence, exhaustivité, et concision.

7 Conclusions et travaux futurs

Cet article a présenté l'ontologie maritime eCISE-OWL ainsi que le processus qui a permis de l'obtenir à partir du modèle de données eCISE. Ce processus comporte deux étapes : i) la conversion automatique de fichiers XSD en langage OWL selon une approche qui améliore l'état de l'art, et ii) des efforts de validation manuelle par des experts du domaine. Cette ontologie est la première tentative de valorisation du modèle de données CISE suivant une approche sémantique de telle sorte que le résultat (l'ontologie et le code de transformation) soit rendu public et disponible pour tous.

A l'avenir, nous envisageons de poursuivre ce travail suivant plusieurs pistes. Un premier ensemble de perspectives vise à améliorer le processus, décrit dans cet article, de génération d'ontologie à partie d'une source XSD : améliorer l'extraction des terminologies des sources externes car pour certaines classes d'association et pour certaines valeurs d'énumérations, l'extraction de termes des tableaux présents dans les fichiers sources exige l'utilisation d'outils d'extraction d'information plus sophistiqués que ceux que nous avons retenus ; revoir les divers espaces de nom actuellement présents au sein de l'ontologie, pour n'en définir qu'un seul associé à l'ensemble de l'ontologie, afin que toutes les entités aient des URI déréférençables dans cet espace (cela suppose de gérer des relations dont les identifiants sont identiques dans chacun des espaces actuels).

Un deuxième ensemble d'objectifs vise à rendre l'ontologie

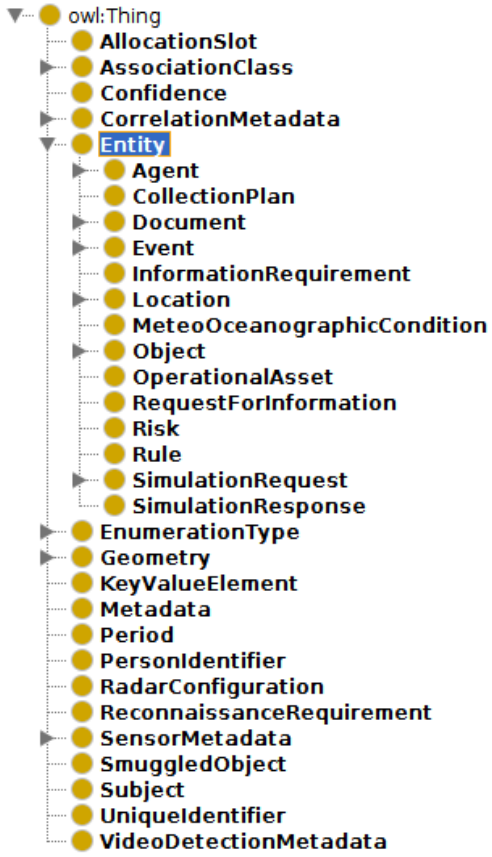


FIGURE 2 – Hiérarchie des concepts principaux d'eCISE-OWL.

15. http://oops.linkeddata.es/catalogue.jsp

Métrique	*eCISE-OWL*	*EUCISE-OWL*
Nombre de classes	268	153
Nombre de propriétés d'objets	297	127
Propriété d'objet - nombre d'axiomes de domaine	336	116
Propriété d'objet - nombre d'axiomes de range	310	116
Nombre de propriétés de données	332	135
Propriété des données - nombre d'axiomes de domaine	350	132
Propriété des données - nombre d'axiomes de range	332	132
Nombre d'individus	16,423	869
Expressivité DL	OWL 2	SHIF(D)
Nombre de triplets	55,322	6,209
Nombre de classes d'association	29	10
Nombre de classes d'énumération	141	

TABLE 2 – Métriques des ontologies eCISE-OWL et EUCISE-OWL [14]

conforme à des principes de partage et de réutilisation. Pour cela, nous allons : générer des alignements entre eCISE-OWL et des ontologies existantes (GeoSPARQL, FOAF, SOSA, etc.) ; rendre l'ontologie entièrement conforme aux principes du FAIR ; gérer les différentes versions des ontologies, issues des différents schémas XSD, y compris la gestion des metadonnées décrivant les ressources non-ontolologiques utilisées comme sources.

Enfin, l'ontologie sera utilisée au sein du système d'information en cours de réalisation dans le projet EFFECTOR, ce qui nécessitera de mettre en place un processus de conversion de messages CISE en messages eCISE en passant par leur représentation en RDF.

Remerciements

Ces travaux sont financés par le programme Horizon 2020 d'innovation et de recherche de l'Union Européenne avec l'accord de financement No. 883374. Ce document ne reflète que la vision des auteurs et la REA (Research Executive Agency) ainsi que la Commission Européenne ne peuvent être tenue responsables pour tout usage de l'information qu'il contient.

Références

[1] Spyros Antonopoulos, Manolis Tsogas, Marios Moutzouris, Antonis Kostaridis, Aggelos Aggelis, and Leonidas Perlepes. Andromeda D.3.1 e-CISE Data Model description. Andromeda project (n. 833881) deliverable, EU, April 2020.

[2] Peb Ruswono Aryan. Converting xml schema to owl in python, 2013.

[3] Nathalie Aussenac-Gilles, Mouna Kamel, Davide Buscaldi, and Catherine Comparot. Construction semi-automatique d'ontologies à partir d'une collection de pages web structurées. In Raphaël Troncy, editor, *IC 2013 : 24es Journées francophones d'Ingénierie des Connaissances (Proceedings of the 24th French Knowledge Engineering Conference), Lille, France, July 1-5, 2013*, pages 165–180, 2013.

[4] I. Bedini, C. Matheus, P. Patel-Schneider, A. Boran, and B. Nguyen. Transforming xml schema to owl using patterns. In *2011 IEEE Fifth International Conference on Semantic Computing*, pages 102–109, 2011.

[5] D. Berger, J. Hermida, F. Oliveri, and G. Pace. The entity service model for cise–service model specifications. Technical report, Joint Research Centre of the European Commission, 2017.

[6] Stefan Brüggemann, Konstantina Bereta, Guohui Xiao, and Manolis Koubarakis. Ontology-based data access for maritime security. In *Proceedings of the 13th International Conference on The Semantic Web. Latest Advances and New Domains - Volume 9678*, page 741–757, Berlin, Heidelberg, 2016. Springer-Verlag.

[7] Diego Calvanese, Benjamin Cogrel, Sarah Komla-Ebri, Roman Kontchakov, Davide Lanti, Martin Rezk, Mariano Rodriguez-Muro, and Guohui Xiao. Ontop : Answering SPARQL queries over relational databases. *Semantic Web*, 8(3) :471–487, 2017.

[8] G.K.D. de Vries, V. Malaisé, M van Someren, P. Adriaans, and A.T. Schreiber. Semi-automatic ontology extension in the maritime domain. In *In Proceedings of the 20th Belgian-Netherlands Conference on Artificial Intelligence (BNAIC 2008)*, 2008. DeVries :BNAIC2008 University of Twente, Enschede, the Netherlands.

[9] EUCISE2020. Eucise2020 : Technical specifications. deliverable 4.3, revision 1, annex b. Technical report, EUCISE Data Model, 2020.

[10] Mokhtaria Hacherouf, S. N. Bahloul, and C. Cruz. Transforming `XML` schemas into `OWL` ontologies using formal concept analysis. *Software & Systems Modeling*, 18 :2093–2110, 2017.

[11] M. Hagaseth, L. Lohrmann, A. Ruiz, F. Oikonomou, D. Roythorne, and S. Rayot. An ontology for digital maritime regulations. *Journal of Maritime Research*, XIII(II) :7–18, 2016.

[12] Aniello Minutolo, Angelo Esposito, Mario Ciampi, Massimo Esposito, and G. Cassetti. An automatic method for deriving OWL ontologies from XML documents. In *2014 Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, Guangdong, China, November 8-10, 2014*, pages 426–431. IEEE Computer Society, 2014.

[13] María Poveda-Villalón, Asunción Gómez-Pérez, and Mari Carmen Suárez-Figueroa. OOPS ! (OntOlogy Pitfall Scanner !) : An On-line Tool for Ontology Evaluation. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 10(2) :7–34, 2014.

[14] Marina Riga, Efstratios Kontopoulos, Konstantinos Ioannidis, Spyridon Kintzios, Stefanos Vrochidis, and Ioannis Kompatsiaris. EUCISE-OWL : an ontology-based representation of the common information sharing environment (CISE) for the maritime domain. *Semantic Web*, 12(4) :603–615, 2021.

[15] Georgios M. Santipantakis, George A. Vouros, Apostolos Glenis, Christos Doulkeridis, and Akrivi Vlachou. The datacron ontology for semantic trajectories. In Eva Blomqvist, Katja Hose, Heiko Paulheim, Agnieszka Ławrynowicz, Fabio Ciravegna, and Olaf Hartig, editors, *The Semantic Web : ESWC 2017 Satellite Events*, pages 26–30, Cham, 2017. Springer International Publishing.

[16] Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, and Mariano Fernández-López. The neon methodology framework : A scenario-based methodology for ontology development. *Appl. Ontology*, 10(2) :107–145, 2015.

[17] Vojtěch Svátek and Ondrej Sváb-Zamazal. Entity naming in semantic web ontologies : Design patterns and empirical observations. In *9th Czecho-Slovak Knowledge Engineering Conference*, 2009.

[18] Arnaud Vandecasteele, Rodolphe Devillers, and Aldo Napoli. A semi-supervised learning framework based on spatio-temporal semantic events for maritime anomaly detection and behavior analysis. In *Coast-GIS 2013 - The 11th International Symposium for GIS and Computer Cartography for Coastal Zone Management*, page 4 pages, Victoria, Canada, June 2013.

[19] Arnaud Vandecasteele and Aldo Napoli. An enhanced spatial reasoning ontology for maritime anomaly detection. In *2012 7th International Conference on System of Systems Engineering (SoSE)*, pages 1–6, 2012.

[20] Arnaud Vandecasteele and Aldo Napoli. Spatial ontologies for detecting abnormal maritime behaviour. In *2012 Oceans - Yeosu*, pages 1–7, 2012.

[21] Boris Villazón-Terrazas and Asunción Gómez-Pérez. Reusing and re-engineering non-ontological resources for building ontologies. In Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, Enrico Motta, and Aldo Gangemi, editors, *Ontology Engineering in a Networked World*, pages 107–145. Springer, 2012.

[22] Diego Vinasco-Alvarez, John Samuel Samuel, Sylvie Servigne, and Gilles Gesquière. From CityGML to OWL. Technical report, LIRIS UMR 5205, September 2020.

[23] Mustafa Yüksel. A semantic interoperability framework for reinforcing post market safety studies. Technical report, Middle East Technical University, 2013.

[24] Dong Xia Zheng and Xue Da Sun. A knowledge acquisition model in maritime domain based on ontology. In *Proceedings IEEE International Conference on Security, Pattern Analysis, and Cybernetics, SPAC 2014, Wuhan, China, October 18-19, 2014*, pages 372–375. IEEE, 2014.

Intégration sémantique de données Raster pour l'observation de la Terre sur des unités territoriales

Ba Huy Tran, Nathalie Aussenac-Gilles, Catherine Comparot, Cassia Trojahn

[1] IRIT, Université de Toulouse, CNRS, UT1, UT2, Toulouse, France

prenom.nom ou prenom.nom-composé@irit.fr

Résumé

*En Observation de la Terre, le format raster, standard de représentation de données et images, convient mal pour caractériser des zones d'intérêt par la seule valeur des pixels. Nous proposons d'intégrer sémantiquement les données raster à d'autres données sur la base de leurs propriétés spatio-temporelles. Ce processus s'appuie sur un modèle sémantique de données qualifiant une zone géographique grâce à des unités territoriales et sur un processus sémantique d'*extraction, transformation et chargement *(ETL) associant données agrégées et zones géographiques. Cet article est un résumé des travaux présentés dans [3].*

Mots-clés

Intégration sémantique de données spatio-temporelles ; Observation de la Terre ; détection de changement.

Abstract

In Earth Observation, the raster format, standard for data and image representation, is not well suited to characterize areas of interest by pixel values alone. We propose to semantically integrate raster data with other data according to their spatio-temporal properties. This process is based on a semantic data model qualifying a geographical area thanks to territorial units and on a semantic process of extraction, transformation and loading *(ETL) associating aggregated data and geographical areas. This paper presents a summary of the work in [3].*

Keywords

Semantic data integration ; spatial and temporal data ; Earth observation ; change detection.

1 Problématique

Depuis 2015, les satellites Sentinel-1 et Sentinel-2 du programme Copernicus [1] ont fourni un grand volume d'images de haute qualité de la Terre (environ 8-10 To de données par jour), offrant aux utilisateurs des données et des métadonnées d'Observation de la Terre (OT) gratuites, fiables et actualisées. Associées au développement d'algorithmes d'apprentissage automatique, ces sources de données ont stimulé le traitement des images et son application dans divers domaines. Elles ont ouvert la voie à de nouvelles applications, de l'agriculture à la gestion des forêts, ou la surveillance de catastrophes naturelles.

Un des formats de données les plus courants pour gérer des images satellite est le format raster. Un raster modélise les phénomènes géographiques comme une surface régulière dans laquelle chaque cellule (ou pixel) est associée à un indicateur (par exemple un indicateur de végétation) ou à une valeur de phénomène selon un codage ou une classification prédéfinie (comme le codage d'un niveau de changement). Ces représentations peuvent être construites automatiquement, y compris par apprentissage automatique. Plusieurs rasters sont disponibles pour la même zone géographique, ce qui permet de surveiller soit le même phénomène à différentes dates, soit des phénomènes différents ; ils peuvent être comparés ou combinés pour en générer un nouveau raster [4]. Cependant, dans une perspective de prise de décision, l'interprétation de leur contenu nécessite des données ou des représentations de connaissances de plus haut niveau associées à des caractéristiques qui donnent un sens à certaines zones d'intérêt sur Terre.

Cet article traite de l'intégration de données calculées à partir de rasters et de données ouvertes sur la base de leurs propriétés spatiales et temporelles afin de qualifier des zones géographiques d'intérêt. Nous utilisons la notion d'*unité territoriale*, définie comme une division d'un territoire plus vaste, selon un critère lié aux activités humaines (administration, droit, agriculture, ...) et normalisée dans des nomenclatures légalement définies. Les *zones d'intérêt* correspondent alors à certaines unités territoriales sélectionnées pour leur pertinence pour une tâche donnée. Elles sont généralement représentées sous forme d'entités géospatiales dans un format vectoriel. Le croisement de données raster (au format matriciel) avec ces zones d'intérêt et même avec d'autres données au format vectoriel n'est donc pas trivial.

Dans la lignée des travaux de [1, 2], nous proposons de représenter ces données au sein de graphes de connaissances, afin d'en faciliter l'intégration, mais aussi l'exploitation, l'interrogation ainsi qu'une restitution intelligible auprès des utilisateurs. Nous sommes intéressés à étudier (i) quel type d'ontologie est nécessaire pour soutenir l'extraction de connaissances à partir de données de raster d'OT et pour décrire de manière homogène les résultats d'analyse de ces données, également fournis au format raster ; (ii) comment rendre accessibles et utilisables des données d'OT riches

1. http://www.copernicus.eu/en

collectes grâce au traitement d'images et d'autres types d'OT ; et (iii) comment améliorer la traçabilité des données au fil de leur traitement (sources de données, calcul des rasters, processus sémantique) pour améliorer la confiance des utilisateurs et l'exploitation des données.
Le projet européen CANDELA [2] vise à créer une plateforme fournissant des modules et des services permettant aux utilisateurs de manipuler, d'explorer et de traiter rapidement les données Copernicus ainsi que de grands ensembles de données ouvertes. Nous contribuons à ce projet en proposant une intégration sémantique des données tirées des images raster et de données ouvertes, et par un module de recherche sémantique sur les données intégrées.

2 Contributions

Les principales contributions de cette étude concernent les composants suivants :
Un modèle sémantique générique qui permet la description sémantique et homogène de données spatio-temporelles pour qualifier des zones prédéfinies et garde la trace de leur provenance. Ce modèle est extensible pour traiter tout type de propriété d'OT observée et a été appliqué à plusieurs cas d'utilisation. Il est composé de plusieurs sous-modèles interconnectés décrivant les différents types de données : *tom*, un modèle d'observation territoriale permettant de représenter les unités territoriales et les observations associées (tirées des fichiers raster) ; *eom*, un modèle d'OT permettant de représenter les métadonnées des images Sentinel ; et *eoam*, un modèle d'analyse d'OT permettant de représenter les rasters produits par le traitement des images ou les vecteurs.
Un processus sémantique configurable et reproductible de type *Extraction, transformation et chargement* (ETL) [3] basé sur le modèle proposé. Nous avons défini un ensemble de fonctions de transformation pour peupler le modèle sémantique avec des données et obtenir une représentation sémantique homogène des données. Ce processus extrait les données des rasters et les agrège avec des données provenant d'autres sources. L'agrégation a lieu sur les zones des unités territoriales. Ensuite, ce processus relie les données extraites aux concepts du modèle sémantique et assigne les données à une unité territoriale.
Un éco-système EO Sentinel qui permet d'exploiter des données matricielles tirées d'images Sentinel (disponibles au format raster), de représenter et de calculer différentes propriétés à partir de ces données puis d'importer d'autres ensembles de données géolocalisées, matricielles ou vectorielles, à partir de sources externes (par exemple, des données sur la couverture du sol).

3 Evaluation

Nous avons évalué notre approche en termes d'adaptabilité du modèle proposé pour répondre à différents cas d'utilisation (surveillance des vignobles et de l'expansion urbaine), l'adaptabilité de la chaîne de traitement, et la valeur ajoutée des ensembles de données générés pour aider à la prise de décision. Nous discutons également de l'évolutivité de l'approche et de la relation entre la résolution de l'image et la taille des unités territoriales de référence.
Les données sémantiques générées pour ces cas d'utilisation sont stockées dans une base de données à laquelle on peut accéder via une interface de recherche sémantique ou un point d'accès SPARQL [4].

4 Conclusions et travaux futurs

L'approche que nous proposons pour intégrer des données calculées à partir de rasters et d'autres données ouvertes permet de qualifier des unités territoriales sur la base de leurs caractéristiques spatiales (vecteurs) et temporelles.
Nous prévoyons d'étendre ce travail pour assurer le passage à l'échelle de l'approche, en exploitant des scénarios de big data pour la gestion des zones Natura 2000 (pertinentes pour l'étude de l'évolution de l'occupation du sol et la détection des changements dans les zones de conservation). Nous prévoyons également de traiter des données provenant de fichiers CSV, en particulier les observations météorologiques de Météo France [5]. Enfin, considérant les cubes de données comme des tableaux de données multidimensionnels fréquemment utilisés pour enregistrer des données géolocalisées, nous envisageons que le processus d'intégration gère ce type de structure, ce qui est rarement le cas dans l'état de l'art actuel.

Remerciements

Ce travail a bénéficié d'une subvention H2020 de la Communauté Européenne pour le projet CANDELA.

Références

[1] L. Ding, G. Xiao, D. Calvanese, and L. Meng. A framework uniting ontology-based geodata integration and geovisual analytics. *ISPRS International Journal of Geo-Information*, 9(9), 2020.

[2] Daniela Espinoza-Molina, Charalampos Nikolaou, Corneliu Octavian Dumitru, Konstantina Bereta, Manolis Koubarakis, Gottfried Schwarz, and Mihai Datcu. Very-High-Resolution SAR Images and Linked Open Data Analytics Based on Ontologies. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(4) :1696 – 1708, 2015.

[3] Ba-Huy Tran, Nathalie Aussenac-Gilles, Catherine Comparot, and Cassia Trojahn. Semantic integration of raster data for earth observation on territorial units. *ISPRS Int. Journal of Geo-Information*, 11(2), 2022.

[4] Jesús Villegas, Hector Sánchez Pastor, Lorena Hernanz, María Checa, and Dumitru Roman. Enabling the use of sentinel-2 and lidar data for common agriculture policy funds assignment. *International Journal of Geo-Information*, 6 :255, 08 2017.

2. http://www.candela-h2020.eu/
3. Accès à l'image docker qui encapsule ce processus :https://hub.docker.com/r/h2020candela/triplification
4. http://melodi.irit.fr/tom/
5. http://www.meteofrance.com/

Session 5 : Explicabilité et interprétabilité dans les graphes de connaissances

Evaluation d'explications pour la prédiction de liens dans les graphes de connaissances par des réseaux convolutifs

Fabien Gandon[1], Nicholas Halliwell[1], Freddy Lecue[1,2]

[1] Inria de l'Université Côte d'Azur, I3S, CNRS
[2] CortAIx, Thales

prénom.nom@inria.fr

Résumé

Nous résumons ici l'article [1] dont l'approche permet de générer un jeu de test comparatif et fournit des métriques pour évaluer les explications de prédictions de liens dans les graphes de connaissances par des réseaux convolutifs pour les graphes relationnels et ceci en présence de plusieurs explications possibles.

Mots-clés

prédiction de liens, IA explicable, graphes de connaissances, réseaux de neurones, évaluation d'explication.

Abstract

We summarize the paper [1] which the approach allows to generate a benchmarks and provides metrics to evaluate explanations of link predictions in knowledge graphs by relational graphs convolutional networks when several possible explanations do exist.

Keywords

link prediction, explainable AI, knowledge graphs, graph neural networks, explanation evaluation.

1 Introduction

Les réseaux convolutifs dédiés aux graphes relationnels (RGCN) [2, 3] sont utilisés sur des graphes de connaissances [4] notamment pour prédire des liens manquants mais malheureusement comme des boîtes noires. Plusieurs méthodes de génération d'explications ont été proposées pour expliquer leurs prédictions. Cependant la performance des méthodes d'explication est difficile à évaluer en absence d'une vérité terrain. De plus, il peut y avoir plusieurs explications pour une même prédiction dans un graphe de connaissances. Jusqu'à présent, il n'existait aucun jeu de données où les observations avaient de multiples explications avec lesquelles se comparer. De plus, il n'existait pas de métrique standard pour comparer les explications prédites par rapport à de multiples explications possible. Dans l'article [1], nous avons proposé une méthode et un jeu de données (FrenchRoyalty-200k), pour évaluer les performances de systèmes d'explication des RGCN sur la tâche de prédiction de liens dans un graphe de connaissances en présence de plusieurs explications possibles. De plus nous avons mené une expérience où des utilisateurs ont évalué chaque type d'explication possible de la vérité terrain en fonction de leur compréhension de l'explication et ceci afin d'affiner l'évaluation de la qualité des explications choisies par un système. A partir de cette vérité terrain, nous proposons l'utilisation de plusieurs métriques utilisant des poids agrégeant les scores des utilisateurs pour chaque explication prédite. Pour valider notre approche, nous nous avons évalué sur ce jeu de données des méthodes d'explication récentes pour la prédiction de liens en utilisant les mesures proposées.

2 De la nécessité d'évaluer les explications des prédictions de liens

Peu d'algorithmes existent pour aider à comprendre les prédictions des RGCN sur un graphe de connaissances. ExplaiNE [2] mesure le changement du score de sa méthode dû à une petite perturbation dans la matrice d'adjacence du graphe pour évaluer l'importance de la présence ou l'absence d'un lien dans la prédiction d'un autre lien. ExplaiNE repose sur l'hypothèse qu'une explication peut être fournie en sélectionnant l'un des voisins de la prédiction. GNNExplainer [3] explique les prédictions en apprenant un masque sur la matrice d'adjacence d'entrée pour y identifier le sous-graphe le plus impactant pour la prédiction. Une faiblesse de ces méthodes est l'évaluation de la qualité de l'explication. Les auteurs d'ExplaiNE reconnaissent eux-mêmes qu'il est difficile de mesurer la qualité de l'explication en l'absence de vérité terrain [2]. ExplaiNE mesure la qualité de ses explications en utilisant la similarité de Jaccard moyenne entre les genres pour un film recommandé donné, et l'ensemble des genres des 5 premières explications sélectionnées. Il s'agit d'une évaluation très limitée qui ne se généralise pas à d'autres tâche ou graphes facilement. De même, GNNExplainer n'a pas été évalué sur la tâche de prédiction de liens explicables sur les graphes de connaissances. Les évaluations sont donc limitées et, a fortiori, ne permettent pas non plus de comparer les méthodes d'explication entre elles.

Dans l'article [5] nous avions commencé par proposer une méthode et des ressources pour évaluer quantitativement et qualitativement les méthodes d'explication sur la tâche de

prédiction de liens dans un graphe de connaissances à partir de données du Web sémantique et dans le cas d'explications uniques pour chaque prédiction. Dans l'article [1] nous avons consolidé et étendu ces résultats en fournissant un autre jeu de données incluant de multiples explications possibles pour une même prédiction et en introduisant des métriques permettant l'évaluation et la comparaison des méthodes de prédiction.

3 Des traces d'inférences notées par les utilisateurs comme explications

Dans un graphe de connaissances, la sémantique formelle disponible nous permet de proposer comme vérité terrain pour des explications de prédiction de liens la justification implication logique de ce lien. Grâce à un raisonneur sémantique open-source avec des capacités de traçage de règles [6] nous avons généré automatiquement des explications pour des règles dérivant de nouveaux liens. Ce traçage identifie les liens qui ont causé la génération d'un autre lien que nous pouvons soumettre à des méthodes qui tenteront à leur tour de le prédire et de s'en expliquer. Cette approche générique de génération d'explications de la vérité du terrain peut être appliquée à de nombreux graphes de connaissances et à de nombreux ensembles de règles. De plus en multipliant les règles nous pouvons générer plusieurs explications possibles pour un lien. Certaines explications peuvent être plus faciles à comprendre que d'autres et l'évaluation d'une méthode devrait prendre cela en compte. Nous avons mené une expérience auprès d'utilisateurs pour noter chaque type d'explication possible. Cela nous permet de distinguer les explications qui sont intuitives de celles qui ne le sont pas, sans nous appuyer sur des hypothèses préalables. Au total, 42 utilisateurs ont répondu, de 11 nationalités différentes, issus de milieux informatiques et non informatiques. Nous avons normalisé les scores moyens entre 0 et 1 pour chaque explication possible, et les avons arrondis au dixième le plus proche pour en faire des poids utilisables dans l'évaluation des explications choisies par un système. Toutes les ressources utilisées et produites sont disponibles en ligne, y compris le lien de téléchargement du raisonneur, le code et les jeux de données [1].

4 Métriques, évaluation et résultats

Nous avons proposé d'évaluer les méthodes d'explication en adaptant la précision et le rappel généralisés [6] proposés à l'origine pour la recherche de documents. Nous définissons aussi la mesure F_1 généralisée, comme la moyenne harmonique entre précision et rappel généralisés et nous proposons la métrique max-Jaccard pour identifier quelle explication a le plus de points communs avec une explication prédite. La métrique max-Jaccard mesure à quel point une méthode d'explication est capable de prédire avec précision une des explications possibles. La précision et le rappel généralisés intègrent à quel point l'explication prédite a reçu un score élevé de la part les utilisateurs. Ces deux mesures empêchent une méthode d'explication de prédire uniquement des explications peu intuitives tout en recevant un score élevé. La mesure F_1 généralisée fournit une vue d'ensemble de la performance.

A partir des traces de règles appliquées à un jeu de données issu de DBpedia nous avons construit un graphe de connaissances de plus de 200 000 triplets avec différentes explications possibles et leurs scores. En utilisant les mesures introduites, nous obtenons un jeu de test permettant d'évaluer et comparer quantitativement différentes méthodes d'explication. Nous montrons que les méthodes d'explication n'essaient pas toujours de prédire la meilleure explication possible (i.e. celle avec le meilleur score des utilisateurs) et qu'elles tentent parfois de prédire des explications avec un score inférieur et n'y réussissent que partiellement. Le graphe de connaissance et son schéma nous permettent aussi d'effectuer une analyse d'erreur sur les prédictions les plus fréquentes et une comparaison des comportements en fonction des caractéristiques des liens (ex. symétriques, inverses, etc.).

Au final, la méthode introduite dans l'article, son jeu de données et ses métriques permettent aux chercheurs de développer de nouvelles méthodes d'explication et d'évaluer quantitativement et qualitativement leurs résultats d'une manière qui leur était auparavant impossible.

1. https ://github.com/halliwelln/multiple-explanations/

Remerciements

Ce travail est soutenu par le 3IA Côte d'Azur - ANR-19-P3IA-0002

Références

[1] N. Halliwell, F. Gandon, F. Lecue, *User Scored Evaluation of Non-Unique Explanations for Relational Graph Convolutional Network Link Prediction on Knowledge Graphs*, InProceedings of the 11th on Knowledge Capture Conference 2021 Dec 2 (pp. 57-64).

[2] B. Kang, J. Lijffijt, T. De Bie, *ExplaiNE : An Approach for Explaining Network Embedding-based Link Predictions*, CoRR abs/1904.12694 (2019).

[3] Z. Ying, D. Bourgeois, J. You, M. Zitnik, J. Leskovec, *GNNExplainer : Generating Explanations for Graph Neural Networks*, In Advances in Neural Information Processing Systems, 2019.

[4] F. Gandon, *Dessine-moi un graphe de connaissances !*, Binaire, 5 Oct. 2021.

[5] N. Halliwell, F. Gandon, F. Lecue, *Linked Data Ground Truth for Quantitative and Qualitative Evaluation of Explanations for Relational Graph Convolutional Network Link Prediction on Knowledge Graphs*, WI-IAT 2021 - 20th IEEE/WIC/ACM Int. Conference on Web Intelligence and Intelligent Agent Technology, 2021.

[6] O Corby, A Gaignard, C Faron Zucker, J Montagnat. *KGRAM Versatile Inference and Query Engine for the Web of Linked Data*, In IEEE/WIC/ACM Int. Conference on Web Intelligence, 2012.

Utiliser les connaissances du sens commun pour la découverte des topics intérpretables

I. Harrando[1], R. Troncy[1]

[1] EURECOM, Sophia Antipolis

mél

Résumé

Les approches traditionnelles de modélisation de sujets (Topic Modeling) s'appuient généralement sur des statistiques de cooccurrence entre termes et documents pour trouver des sujets latents dans une collection de documents. Cependant, le fait de s'appuyer uniquement sur ces statistiques peut donner des résultats incohérents ou difficiles à interpréter pour les utilisateurs finaux dans de nombreuses applications où l'intérêt réside dans l'interprétation des sujets résultants (e.g. l'étiquetage de documents, la comparaison de corpus, orienter l'exploration du contenu..). Nous proposons de tirer parti des connaissances externes de sens commun, c'est-à-dire des informations du monde réel au-delà de la cooccurrence des mots, pour trouver des topics plus cohérents et plus facilement interprétables par les humains. Nous présentons le "Common Sense Topic Model" (CSTM), une approche nouvelle et efficace qui augmente le clustering avec des connaissances extraites du graphe de connaissances ConceptNet. Nous évaluons cette approche sur plusieurs jeux de données en comparaison avec des modèles couramment utilisés, en utilisant une évaluation automatique et humaine, et nous montrons comment elle montre une corrélation supérieure au jugement humain.
Cet article a été déjà publié à K-CAP 2021[4].

Mots-clés

Exemple type, format, modèle.

1 Introduction

Le Topic modeling (modélisation de sujets) est une technique de fouille de textes qui est largement utilisée pour de nombreuses applications, à la fois pour d'autres tâches dites "downstream" du TAL (e.g. la similarité de textes), mais aussi comme un outil pour explorer, visualiser et interpréter le contenu de grandes collections de textes. Alors que la première application peut être évaluée et améliorée en mesurant quantitativement la performance sur la tâche elle-même, il est plus difficile de saisir la capacité d'un algorithme de topic modeling à générer des résultats compréhensibles et utiles pour un utilisateur humain. Plusieurs efforts de recherche antérieurs [1, 2] ont mis en évidence la divergence entre la plupart des mesures d'évaluation automatiques (largement utilisées dans la littérature) et le jugement humain, car ces modèles ont tendance à optimiser pour des objectifs numériques qui s'alignent ou se corrèlent rarement bien avec ce que les humains considèrent comme des "sujets" (topics).

La plupart des approches de modélisation des sujets se concentrent sur les statistiques de co-occurrence des mots comme signal principal pour détecter les relations sémantiques latentes entre eux – une idée qui remonte aux années 50 (*"Vous connaîtrez un mot par la compagnie qu'il garde"*[3]). Cela les rend intrinsèquement incapables de capturer les relations entre les mots qui ne sont pas explicitement présents dans les données d'apprentissage. De nombreux travaux ont été réalisés pour explorer la possibilité d'injecter des connaissances externes (généralement spécifiques à un domaine) dans la tâche de modélisation de sujets. Pourtant, bien que l'utilisation du sens commun a été explorée pour la classification des topics [5], aucune tentative d'incorporation de connaissances générales humaines (ou *sens commun*) dans le processus de modélisation de sujets n'a été proposée pour combler le fossé entre l'optimisation basée sur les statistiques et le jugement humain. Nous proposons une méthode qui combine les connaissances dans un graphe de connaissances dit de sens commun [7] avec un algorithme de clustering pour produire des sujets qui sont plus corrélés avec le jugement humain de la cohérence tout en s'adaptant sans problème à de grands ensembles de données.

2 Approche

Comme dans les travaux précédents [6], nous abordons la tâche de modélisation des sujets comme un *problème de clustering de documents*, c'est-à-dire que nous générons des représentations vectorielles pour tous les documents du corpus étudié que nous appelons *Sac de mots enrichi en sens commun (Common-sense enriched Bag-of-words)*, puis nous exécutons un algorithme de clustering pour trouver N groupes cohérents (N étant le nombre de sujets) qui représentent nos sujets. On appelle ce modèle *CSTM* (Common-Sense Topic Model). La figure 1 représente illuste cette approche.

Common-sense Enriched Bag of Words (CS-BoW). Inspirés par des méthodes issues de la littérature sur l'expansion de requêtes, nous proposons d'enrichir la représentation "sac de mots" des documents, avec des termes connexes issus du graphe de connaissances ConceptNet.

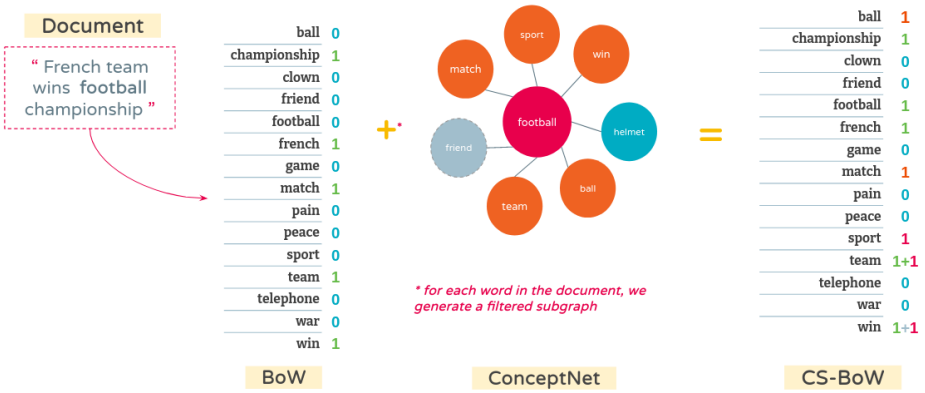


FIGURE 1 – Illustration of the process of creating the common sense-enhanced document representation using ConceptNet. We note that we filter out words that do no appear in the vocabulary (friend) as well as words with low similarity (helmet).

L'avantage d'utiliser ConceptNet est qu'il est principalement peuplé par la relation "Relié à", qui implique une relation topique entre les termes. Concrètement, pour chaque mot du document, nous interrogeons ConceptNet pour récupérer tous les termes qui lui sont directement liés (à un saut de puce sur le graphe), et nous les ajoutons au document. Par exemple, un document qui mentionne le mot "caméra" sera automatiquement enrichi avec les mots "photo", "objectif", etc. La représentation du document est alors construite comme un sac de mots contenant tous les mots originaux du document, en plus de tous les mots qui leur sont liés dans ConceptNet.

Clustering. Il existe une littérature riche et variée sur la tâche de clustering. Dans un souci de simplicité, nous choisissons *K-Means*, un algorithme de clustering couramment utilisé, rapide et capable de traiter des ensembles de données plus importants à l'aide de l'implémentation hautement optimisée *FAISS* [1], et nous l'exécutons sur les représentations CS-BoW des documents du corpus. Pour générer les mots clés du sujet, nous considérons les vecteurs centroïdes générés par K-Means et choisissons les composantes (correspondant aux mots sur la représentation CS-BoW) avec les plus grands coefficients pour représenter le sujet.

3 Evaluation

On propose de faire l'évaluation de notre modèle (en le comparant avec deux autres baselines) en deux étapes : évaluation automatique (quantitative) avec les métriques classiques utilisées dans la litérature, et puis une évaluation qualitative faite par 12 personnes qui parlent l'anglais couramment. On leur demande d'effectuer trois tâches pour évaluer les topics résultants (Intrusion de mots, labélisation des sujets, classification des sujets).

On observe globalement que malgré le fait que CSTM n'est pas toujours le meilleur au niveau des métriques automatiques, il dépasse de manière considérablement les autres modèles sur les tâches de l'évaluation humaine, ce qui montre que les sujets générés par CSTM sont plus facilement intérprétés par les humains.

1. https://github.com/facebookresearch/faiss/

Remerciements

Ce travail a été partiellement soutenu par le programme de recherche et d'innovation Horizon 2020 de l'Union européenne dans le cadre du projet Odeuropa (accord de subvention n° 101004469), par CHIST-ERA dans le cadre du projet CIMPLE (CHIST-ERA-19-XAI-003) et par raisin.ai dans le cadre du projet MyLittleEngine.

Références

[1] Jonathan CHANG et al. "Reading Tea Leaves : How Humans Interpret Topic Models". In : NIPS'09. Vancouver, Canada, 2009.

[2] Caitlin DOOGAN et Wray BUNTINE. "Topic Model or Topic Twaddle ? Re-evaluating Semantic Interpretability Measures". In : *NAACL '21*. Juin 2021.

[3] Adriana FERRUGENTO et al. *A synopsis of linguistic theory 1930-1955*. 1957.

[4] Ismail HARRANDO et Raphaël TRONCY. "Discovering Interpretable Topics by Leveraging Common Sense Knowledge". In : K-CAP '21. USA, 2021.

[5] Ismail HARRANDO et Raphaël TRONCY. "Explainable Zero-Shot Topic Extraction Using a Common-Sense Knowledge Graph". In : *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Dagstuhl, Germany.

[6] Suzanna SIA, Ayush DALMIA et Sabrina J. MIELKE. "Tired of Topic Models ? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too !" In : *EMNLP '20*. Online : ACL, nov. 2020.

[7] Robyn SPEER, Joshua CHIN et Catherine HAVASI. "ConceptNet 5.5 : An Open Multilingual Graph of General Knowledge". In : 2017, p. 4444-4451.

Stunning Doodle: un outil pour la visualisation et l'analyse conjointe de graphes de connaissances et leurs plongements

A. Ettorre, A. Bobasheva, F. Michel, C. Faron
Université Côte d'Azur, CNRS, Inria, I3S, France
{aettorre, abobasheva, fmichel, faron}@i3s.unice.fr

Résumé

Ces dernières années, l'utilisation croissante des graphes de connaissances dans des domaines nouveaux et variés nécessite de rendre ces ressources accessibles et compréhensibles par des utilisateurs aux profils de plus en plus divers. De plus, l'application de méthodes d'apprentissage automatique sur des plongements de graphes de connaissances renforce encore la visibilité de ce type de représentation, mais soulève un nouveau problème de compréhension et d'interprétabilité de ces plongements. Dans ce travail, nous montrons comment des techniques de visualisation peuvent être utilisées pour explorer et interpréter conjointement les graphes de connaissances et les plongements de graphes.

Mots-clés

Plongements de graphes de connaissances, Visualisation.

Abstract

In recent years, the growing application of Knowledge Graphs to new and diverse domains has created the need to make these resources accessible and understandable by users with increasingly diverse backgrounds. Moreover, the emerging use of Knowledge Graph Embeddings as input features of Machine Learning methods has given even more visibility to this kind of representation, but raising the new issue of understandability and interpretability of such embeddings. In this paper, we show how visualization techniques can be used to jointly explore and interpret both Knowledge Graphs and Graph Embeddings.

Keywords

Knowledge Graph Embeddings, Visualization.

1 Introduction

Au cours de la dernière décennie, l'adoption des graphes de connaissances (KGs) dans de multiples domaines n'a cessé de croître de sorte que de plus en plus de projets s'appuient sur ce type de représentation pour stocker leurs données en en conservant toute la richesse sémantique. L'une des raisons de leur succès croissant est la possibilité de leur appliquer des méthodes d'apprentissage automatique en utilisant une représentation à faible dimension de ces KGs, obtenue par le processus de plongement de graphes. Néanmoins, il n'est pas aisé de comprendre les informations capturées par les plongements de graphes (GEs). En effet, les GEs sont calculés à l'aide de techniques d'apprentissage automatique, des "boîtes noires" qui traduisent chaque élément du graphe en un vecteur de faible dimension. Même si le processus algorithmique de calcul des plongements est bien compris, une relation entre les caractéristiques et le rôle de l'élément du graphe et sa représentation vectorielle dans l'espace de plongement ne peut être établie avec certitude. En d'autres termes, il n'est pas aisé de répondre à certaines questions, notamment :

— Que représentent mes plongements ?
— Comment sont-ils liés à la structure et à la sémantique de mon KG ?
— Comment puis-je améliorer mes plongements pour qu'ils soient mieux adaptés à mon application downstream ?

Récemment, plusieurs efforts de recherche ont été faits dans cette direction pour commencer à donner du sens aux informations capturées par les GEs. Certaines approches mettent en œuvre des stratégies d'explication pour des modèles de plongement spécifiques [3] ; tandis que d'autres proposent des méthodes pour vérifier si un élément de connaissance spécifique représenté dans un KG est réellement encodé et capturé par ses GEs [1].

Dans [2], nous avons abordé cette question d'un point de vue différent. Nous pensons que l'information portée par les GEs pourrait être explorée et dévoilée grâce à l'utilisation de techniques de visualisation qui favoriseraient la découverte de la relation logique entre le graphe et ses plongements. Notre objectif est donc de fournir un outil de visualisation permettant l'analyse et le décodage des informations capturées par les KGEs en dévoilant la relation entre, d'une part, la structure et la sémantique du KG et, d'autre part, les KGEs générés à partir de celui-ci. Nous présentons *Stunning Doodle* [2], un outil conçu pour la visualisation de KGs basés sur RDF et de leur GEs. *Stunning Doodle* fournit une visualisation du graphe à analyser, offrant un aperçu riche de la structure et de la sémantique des données. Cette visualisation est ensuite enrichie en connectant les sommets du graphe avec les GEs correspondants à analyser.

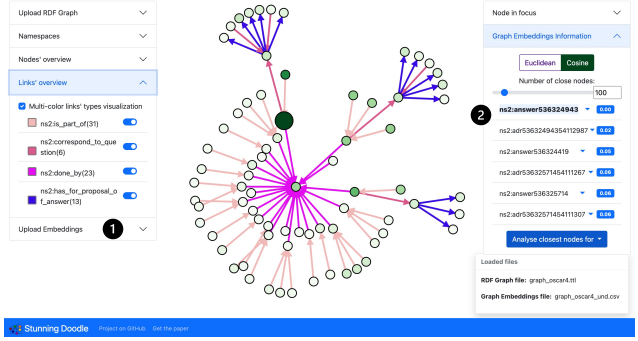


FIGURE 1 – *Stunning Doodle* montrant sommets proches dans l'espace de plongement.

2 Analyse des GEs avec Stunning Doodle

Stunning Doodle a été développé pour combler un manque dans le domaine de l'analyse visuelle des KGEs. Son objectif principal est de fournir aux utilisateurs une visualisation avancée de graphes RDF, enrichie d'informations extraites des GEs générés à partir de ces graphes. Premièrement, *Stunning Doodle* permet aux utilisateurs de visualiser et de naviguer dans les graphes RDF grâce à son système d'exploration de graphe et à l'utilisation de fonctionnalités avancées de filtrage et de personnalisation basées sur la sémantique des sommets et des arêtes.

La fonctionnalité clé fournie par *Stunning Doodle* est la possibilité de visualiser conjointement les KGEs et le KG à partir duquel ils sont générés. Plus précisément, *Stunning Doodle* permet de visualiser, pour chaque sommet, ses sommets les plus proches dans l'espace de plongement, selon la distance euclidienne ou la distance cosinus. La figure 1 montre un exemple d'analyse des KGEs. Après avoir téléchargé un fichier stockant les GEs calculés à partir du KG (1) visualisé, l'utilisateur peut sélectionner un sommet d'intérêt et visualiser ses plus proches voisins dans l'espace de plongement. Les sommets les plus proches sont affichés avec un gradient de couleur qui représente leur distance par rapport au sommet dont le plongement est analysé ; les sommets plus sombres sont plus proches (dans l'espace de plongement) du sommet sélectionné tandis que les sommets plus clairs sont plus éloignés. Si une relation entre un couple de sommets visualisés existe dans le KG, alors l'arête correspondante est directement affichée dans le graphique selon les paramètres de personnalisation sélectionnés. La liste des sommets les plus proches avec leur distance est affichée dans le menu "Graph Embeddings Information" sur le côté droit de la page (2). Outre cette liste, des options supplémentaires permettent à l'utilisateur de choisir la métrique de distance souhaitée (euclidienne ou cosinus dans l'implémentation actuelle) et de personnaliser le nombre de sommets les plus proches à afficher. L'exemple de la figure 1 montre les 100 sommets les plus proches dans l'espace de plongement du sommet `ns1:answer536324943`, selon la distance cosinus. Le sommet actuellement sélectionné (de plus grande taille) est ici le sommet à partir duquel les distances sont calculées. Il est facilement reconnaissable sur la visualisation par sa couleur sombre due au fait que sa distance par rapport à lui-même est nulle. Son URI est donc le premier élément de la liste des sommets les plus proches. Les différentes nuances de vert de chaque sommet mettent clairement en évidence les sommets les plus proches parmi les 100 visualisés, tandis que les liens montrent comment ils sont connectés dans le KG. Tout en visualisant les voisins d'un sommet dans l'espace de plongement, il est toujours possible d'accéder aux fonctionnalités de navigation dans le KG grâce aux boutons de la section "Node in focus". Par conséquent, tout sommet affiché peut être développé pour afficher les sommets qui lui sont liés dans le KG, même s'ils ne sont pas proches dans l'espace de plongement. Naturellement, tout nouveau sommet peut être développé à son tour pour visualiser la partie souhaitée du KG. Toutes les arêtes et les sommets affichés dont les plongements ne sont pas proches du sommet initial seront visualisés en fonction des options de personnalisation sélectionnées dans les sections "Nodes' overview" et "Links' overview".

En résumé, *Stunning Doodle* permet à l'utilisateur de comprendre en un coup d'œil quels sommets d'un KG sont considérés comme similaires dans l'espace de plongement, tout en gardant la trace de leurs connexions dans le KG. Cela permet d'avoir un aperçu immédiat des informations capturées par les KGEs, par exemple quels prédicats ont le plus grand impact ou quels modèles de connectivité sont les plus pris en compte pendant le processus de plongement.

3 Conclusion

Stunning Doodle est une première étape pour combler le manque dans le domaine de l'analyse visuelle des KGEs. Cet outil de visualisation permet de trouver un lien entre le contenu et la structure de n'importe quel KG et ses plongements. Nous avons implémenté un ensemble de fonctionnalités pour faciliter l'exploration et la compréhension de n'importe quel KG et pour analyser les KGEs en connectant les deux, et en donnant du sens aux informations capturées par les KGEs.

Références

[1] Antonia Ettorre, Anna Bobasheva, Catherine Faron, and Franck Michel. A systematic approach to identify the information captured by Knowledge Graph Embeddings. In *IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 2021.

[2] Antonia Ettorre, Anna Bobasheva, Franck Michel, and Catherine Faron. Stunning Doodle : a Tool for Joint Visualization and Analysis of Knowledge Graphs and Graph Embeddings. In *ESWC 2022*, 2022.

[3] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNexplainer : Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32 :9240, 2019.

Session 6 : Apprentissage automatique, ontologies et graphes de connaissances

Découverte de règles causales dans les graphes de connaissances à l'aide de plongements dans les graphes

L. Simonne[1], N. Pernelle[2], F. Saïs[1], R. Thomopoulos[3]

[1] LISN, CNRS (UMR 9015), Université Paris-Saclay
[2] LIPN, CNRS (UMR 7030), Université Sorbonne Paris Nord
[3] INRAE (UMR IATE)

simonne@lisn.fr

Résumé

La découverte de relations causales est l'objectif de nombreuses expériences. Lorsque des données observationnelles sont disponibles, l'utilisation du cadre d'étude des résultats potentiels est un des standards pour découvrir de telles relations. Dans cet article, nous nous plaçons dans ce cadre afin de découvrir des règles causales au sein de graphes de connaissances (KGs). Ces règles expriment que des différences de traitements conduisent à des différences de valeur pour une caractéristique étudiée. Cependant, ce cadre repose sur la similarité entre instances, et sa quantification dans un KG n'est pas triviale, notamment parce que leurs descriptions peuvent être incomplètes et erronées. Nous proposons une nouvelle méthode de découverte de règles causales qui exploite un appariement basé sur les plongements de graphes de connaissances. Les expérimentations menées sur deux KG de domaines différents ont montré la capacité de notre approche à découvrir des règles qui expliquent plus de différences dans la caractéristique étudiée que les approches existantes, et qui est plus robuste en cas d'information incomplète.

Mots-clés

Règles causales, Plongements de graphes, Explicabilité, Graphes de connaissances

Abstract

Discovering causal relationships between different observations is the goal of many experiments in science. When observational data are available, the potential outcome framework is a well-used framework for discovering such relationships. In this paper, we place ourselves in this framework to discover causal rules in Knowledge Graphs (KGs) that express that differences in treatments lead to differences in a studied characteristic. However, quantifying the similarity between individuals represented in a knowledge graph is challenging, especially because their descriptions can be incomplete and erroneous. We propose a new approach based on knowledge graph embeddings to discover causal rules in KGs. The experiments that we conducted on two KGs, including a scientific knowledge graph, showed that our approach is able to discover rules that explain much more differences in the studied characteristic than existing state of the art approaches.

Keywords

Causal Rules, Graph Embeddings, Explainability, Knowledge Graphs

1 Introduction

L'adoption des technologies du web sémantique pour la représentation des données et des connaissances est en plein essor. Cela a conduit à la disponibilité de nombreux ensembles de données représentés sous forme de graphes de connaissances [14] où les données sont représentées en RDF et les connaissances du domaine sous forme d'ontologie exprimée en OWL ou RDFS. Ces graphes de connaissances peuvent contenir des données et des connaissances de différents domaines comme DBPedia et Wikidata ou plus spécifiques à un domaine comme MusicBrainz ou Bio2RDF. Ces KG peuvent être exploités pour découvrir de nouvelles connaissances telles que des règles d'associations (i.e. $hasChild(x, y) \wedge citizenOf(x, z) \Rightarrow citizenOf(y, z)$). Ces associations peuvent être utiles pour prédire de nouveaux faits, détecter des erreurs, mais elles expriment rarement des relations causales.

Une *relation causale* décrit la relation entre deux variables, où une variable nommée *traitement* a un *effet* sur une variable nommée *résultat*. Les relations causales sont intéressantes dans de nombreux domaines, comme par exemple en santé pour déterminer si un médicament traite ou non une maladie, ou en politiques publiques pour comprendre si une nouvelle loi a eu un impact attendu ou non. Il existe différents cadres pour étudier le problème de la découverte de relations causales à partir de données d'observation tabulaires, tels que le *modèle causal structurel* [15] ou le *cadre des résultats potentiels* [18]. Dans ce dernier cas, l'effet d'un traitement peut être estimé en comparant des instances similaires qui diffèrent sur le traitement. Bien que de nombreuses approches existent pour découvrir les relations causales dans les données tabulaires, seules quelques approches [11, 20] se concentrent sur la découverte des effets d'un traitement dans les KG. L'approche proposée dans [20], permet de découvrir des règles causales différentielles

qui expriment qu'une différence de valeurs sur le traitement explique une différence de valeurs sur le résultat. Cette approche, fondée sur le cadre des résultats potentiels, s'appuie sur une étape d'appariement tronqué fondé sur le regroupement de certains chemins de propriétés. Les règles découvertes sont très expressives, car elles représentent explicitement le sous-ensemble de classes d'instances pour lesquelles la règle est valide. Un processus de généralisation des règles est effectué, mais il conduit à très peu de règles générales, en raison de fortes contraintes sur la complétude des données. Ainsi, seule une faible partie des différences de résultats peut être expliquée à l'aide des règles.

Dans cet article, nous présentons une nouvelle approche hybride appelée Dicare-E combinant les plongements de graphes et les techniques de fouille de règles symboliques pour découvrir des règles différentielles causales dans les graphes de connaissances. Ce type de règle permet à la fois d'expliquer des différences de valeur sur une propriété étudiée comme la croissance de la taille d'une tumeur, ou le taux de réinsertion des chômeurs, et d'en tirer des conclusions pour effectuer des décisions. Nous adaptons le cadre des résultats potentiels pour pouvoir découvrir de telles règles. Pour apparier des instances, nous exploitons des plongements pré-entraînés et mesurons la similarité de deux vecteurs intégrés en observant la similarité des prédictions utilisant ces vecteurs. Ainsi, la méthode d'appariement est plus robuste aux données erronées et incomplètes. En outre, elle permet d'obtenir des règles plus générales qui peuvent être appliquées à un plus grand nombre d'instances. Nos contributions sont (i) la définition d'une nouvelle méthode d'appariement basée sur les plongements de graphes, (ii) un algorithme s'appuyant sur ces appariements pour découvrir des règles exprimant des effets causaux de traitements et (iii) une évaluation expérimentale sur un jeu de données réel qui montre l'efficacité de l'approche développée en comparaison avec l'état de l'art.

Dans la section 2, nous présentons les travaux antérieurs sur la découverte de la causalité dans les KG et introduisons le problème fondamental de l'inférence causale et la définition de l'appariement d'instances. Ensuite, dans la section 3, nous présentons l'énoncé du problème que nous abordons dans ce travail. La section 4 explique en détail comment nous utilisons les vecteurs issus des plongements pour calculer la similarité entre deux instances. L'algorithme est présenté dans la section 5. Enfin, dans la section 6, nous présentons les expériences et les résultats obtenus sur un graphe de connaissances scientifiques.

2 Travaux antérieurs

Dans cette section, nous présentons les travaux antérieurs qui traitent de la découverte causale dans les données tabulaires et dans les graphes de connaissances, ainsi que les principales méthodes de comparaison d'instances.

Causalité dans les données tabulaires. La découverte de relations causales est étudiée depuis des décennies et de nombreuses approches ont été définies. Certaines approches sont basées sur le modèle causal structurel, introduit par Pearl [15], où les modèles visent à décrire les systèmes avec des modèles graphiques. D'autres utilisent les réseaux bayésiens [12] qui représentent des ensembles de covariables avec des modèles probabilistes graphiques, qui peuvent montrer des liens causaux sous certaines hypothèses. Comme les relations causales peuvent être difficiles à extraire dans des systèmes décrivant de nombreuses variables [2], certains travaux visent à déterminer les structures causales locales.

L'étude de la découverte de relations causales peut également être introduite par le biais du problème fondamental de l'inférence causale [18]. Étant donné $Y_i(T)$ le résultat de l'individu i lors de l'étude du traitement T, l'effet du traitement individuel est défini comme $TE_i = Y_i(1) - Y_i(0)$, c'est-à-dire la différence entre le résultat de i avec le traitement et le résultat de i sans le traitement. Cependant, le problème de l'inférence causale réside dans le fait que TE_i ne peut pas être calculé, car pour un individu donné i, si $Y_i(1)$ est observé, le contrefactuel $Y_i(0)$ ne peut pas être observé. Le standard pour trouver les effets du traitement est de planifier une expérience avec une attribution aléatoire du traitement. De telles expériences sont difficiles à planifier pour des raisons d'éthique et de coût. Il n'est par exemple pas possible de forcer les gens à commencer à fumer. Par conséquent, la plupart des études causales sont menées en exploitant des données observationnelles, où l'attribution du traitement ne se fait pas de manière aléatoire. Dans ces approches, deux ensembles d'individus sont créés : l'ensemble *traité* des individus $(T = 1)$, et l'ensemble *contrôle* des individus non traités $(T = 0)$. Comparer naïvement les résultats des deux ensembles introduit un biais de sélection, car la distribution des covariables n'est pas la même dans les deux groupes. Par exemple, en étudiant comme traitement l'âge et en résultat la probabilité qu'une personne ait un cancer, l'ensemble de contrôle pourrait être composé d'une majorité d'hommes, et l'ensemble traité d'une majorité de femmes, ce qui introduirait un biais.

Le cadre d'étude des *résultats potentiels* estime un effet en trouvant un contrefactuel pour les individus traités [18]. Le contrefactuel peut être estimé ou déterminé parmi les individus du *contrôle* [21]. Il permet d'équilibrer la distribution des covariables dans les deux ensembles, de sorte qu'ils partagent une probabilité égale de traitement, et supprime le biais de sélection. Une technique est l'appariement, qui vise à sous-échantillonner l'ensemble de données pour équilibrer la distribution des covariables. Elle consiste, pour un individu traité, à le comparer à un individu contrôle similaire. Dans l'exemple précédent, l'appariement consiste à comparer des paires de personnes qui ont le même sexe, la même taille et le même poids (et autres), mais qui diffèrent sur l'âge de traitement. Bien que cette technique soit populaire, l'appariement devient de plus en plus difficile avec la complexité de la description des individus [1]. Plus le nombre d'attributs décrivant les individus est élevé, plus il est difficile de trouver des individus qui partagent les mêmes valeurs sur un ensemble d'attributs. Des alternatives ont donc été proposées. L'appariement tronqué consiste à relâcher la contrainte d'appariement exact. Elle facilite non

seulement le processus d'appariement, mais minimise également le biais dans les effets découverts [7]. L'appariement par score de propension [17] est une autre technique d'appariement reconnue qui repose sur l'utilisation d'un modèle de classification. Ce modèle produit pour chaque instance un score représentant la probabilité que l'instance soit traitée, et les instances ayant un score similaire sont appariées.

Dans le cadre des graphes de connaissances, l'estimation du contrefactuel doit être adaptée pour prendre en compte les descriptions RDF des individus qui peuvent être encore plus complexes que les données tabulaires.

Causalité dans les graphes de connaissances. À notre connaissance, seules trois approches se sont concentrées sur l'exploration des relations causales dans les KG. Dans [11], les auteurs proposent de transformer les données du KG en un schéma relationnel en utilisant des connaissances expertes et d'apprendre un modèle relationnel probabiliste. Cependant, les résultats représentent la distribution conjointe entre les variables et n'indiquent pas nécessairement les relations causales. De plus, comme elle repose sur des réseaux bayésiens, l'approche ne peut pas prendre en compte de grands graphes de connaissances et elle a été évaluée sur un petit ensemble de données ($\approx$ 7k triplets). [20] découvre des règles différentielles causales expressives, qui expriment un effet de traitement pour un sous-ensemble d'instances décrites par un motif de graphe nommé strate. De telles règles sont intéressantes car elles peuvent montrer des effets locaux. Cependant, une règle définie sur une strate donnée est sélectionnée si sa validité est vérifiée sur toutes les strates les plus spécifiques. Une telle contrainte stricte empêche de découvrir des règles génériques dans le cas de strates spécifiques avec trop peu d'instances correspondantes. Enfin, [6] exploite les plongements de graphes afin de découvrir de nouvelles hypothèses scientifiques à partir d'hypothèses recueillies sur un ensemble de papiers scientifiques. Bien qu'il ne soit pas fait mention d'un cadre d'étude de la causalité, cette approche repose sur l'utilisation de plongements d'hypothèses et de papiers pour découvrir de nouvelles hypothèses, et peut s'apparenter au cadre d'étude des résultats potentiels.

Appariement d'instances dans les graphes de connaissances. Il existe de nombreuses approches de liaison de données visant à découvrir les liens d'identité représentés par le prédicat *owl :sameAs* dans les graphes de connaissances (voir [13] pour un aperçu). Cependant, nous ne sommes pas intéressés par les liens d'identité puisque notre objectif est de comparer des entités distinctes. Ces entités devraient différer sur le traitement et sur le résultat, tout en étant très similaires sur le reste de la description RDF. Des prédicats moins stricts comme *identiConTo* [16] ont été définis pour exprimer la relation d'identité entre des entités limitées à un contexte conceptuel donné (c'est-à-dire une sous-partie de l'ontologie), mais cette approche ne fournit pas de similarité quantifiée et n'est pas adaptée aux données incomplètes. Dans des approches récentes, [19] fait de l'appariement strict. Cette approche est utilisable sur des graphes ayant des schémas simples mais n'est pas applicable lorsque le schéma est complexe, l'appariement devenant rare voire impossible. [20] propose un appariement tronqué pour notamment traiter l'incomplétude des données en utilisant des propriétés abstraites qui sont dérivées de clusters de propriétés obtenus grâce à leur co-occurrence. Cependant, le processus de clustering doit être guidé par un expert du domaine lorsque le nombre de propriétés est élevé.

La recherche par similarité a également été étudiée pour la relaxation de requêtes [4]. L'utilisation de telles approches dépend cependant de la définition d'un espace de recherche de requêtes, ce qui n'est pas trivial. La similarité entre les instances peut également être quantifiée comme dans [3], où les auteurs proposent une similarité entre les clauses de Horn qui est basée sur des prédicats et des arguments (non-)partagés. Cette approche repose sur des descriptions de logique de premier ordre complètes et n'est donc pas conçue pour traiter des données incomplètes.

Les modèles de plongements présentent des caractéristiques intéressantes pour notre approche. En effet, si ces techniques capturent la sémantique des entités (i) des instances RDF similaires ont des vecteurs similaires dans l'espace de plongements [22] et (ii) les vecteurs similaires dans l'espace de plongements représenteront des instances RDF similaires [10, 8]. Cependant, pour utiliser ces approches afin de déterminer des instances similaires mais non identiques, il est nécessaire de fournir un grand ensemble d'instances similaires et dissemblables, et de fixer un seuil qui peut être utilisé pour décider que deux instances sont suffisamment similaires.

Fouille de règles d'association. La fouille de règles d'association est un élément important de la recherche au sein des bases de connaissances. Ces règles sont composées d'un corps $\vec{B}$ et d'une tête $\vec{H}$, où $\vec{B}$ peut contenir un ou plusieurs atomes et $\vec{H}$ un atome, et expriment le lien $\vec{B} \Rightarrow \vec{H}$. Les méthodes de fouille de règles d'association, telles que [5], sont couramment utilisées afin d'obtenir de nouvelles connaissances ou encore de supprimer des triplets erronés. Bien qu'une association soit représentée dans ces règles, elle n'indique pas nécessairement un lien de causalité. De plus, une règle d'association n'indique pas clairement l'effet d'un traitement, et les algorithmes utilisés pour les déterminer ne prennent pas en compte la similarité des instances, *i.e.* il n'y a pas d'étape de contrôle réalisée. Par exemple, un effet présent dans la tête d'une règle pourrait être du à un traitement non indiqué dans le corps. Ainsi, une telle règle ne peut être utilisée pour expliquer l'effet d'un traitement.

3 Préliminaires et Définitions

Nous cherchons à découvrir des règles causales exprimées en logique du premier ordre sous la forme $\vec{B} \Rightarrow \vec{H}$ où des différences de valeurs dans $\vec{B}$, représentant le *traitement* expliquent des différences de valeurs dans $\vec{H}$ qui représente le *résultat*. Dans ce qui suit, nous présentons les définitions formelles permettant la définition des règles différentielles causales qui nous intéressent ainsi que la définition de la mesure permettant d'évaluer leur qualité.

3.1 Définitions

Graphes de connaissances. Nous considérons un graphe de connaissances KG défini par une paire $(\mathcal{O}, \mathcal{F})$ où $\mathcal{O}$ est une ontologie représentée en OWL composée d'un ensemble de classes et de propriétés. $\mathcal{F}$ est un ensemble de triplets RDF décrivant des instances de classes de $\mathcal{O}$.

Traitement. Dans le cadre des graphes de connaissances, nous considérons des chemins de propriétés P_t sous la forme $P_t : p_1(X, Y_1) \wedge p_2(Y_1, Y_2) \wedge \ldots \wedge p_n(Y_{n-1}, Y_n)$ de longueur maximale l_{tmax}. Les propriétés se trouvant à l'extrémité de ces chemins peuvent avoir des valeurs catégorielles ou numériques. Ces propriétés peuvent être monovaluées (i.e, fonctionnelles) ou multi-valuées. Par abus de langage, nous considérons la (les) valeur(s) d'un chemin de propriété comme faisant référence à la (les) valeur(s) de la propriété à l'extrémité du chemin. Nous considérons que les règles s'appliquent à deux instances (X_1, X_2) d'une classe de $\mathcal{O}$, et qu'un traitement T représente une différence de valeurs de P_t entre X_1 et X_2.
Nous distinguons deux types de traitements : un *traitement catégoriel* T_c impliquant un chemin de propriétés dont l'extrémité est une propriété catégorielle (e.g, littéral, date, valeurs hiérarchisées) et un *traitement numérique* T_n impliquant un chemin de propriétés dont l'extrémité est une propriété numérique (e.g, entier, réel).

Traitement catégoriel. Soient X_1 et X_2 deux instances d'une classe, un chemin de propriétés P_t et deux ensembles de valeurs V_1 et V_2 du chemin P_t pour X_1 et X_2 respectivement. Un traitement catégoriel T_c est défini par :

$$T_c(X_1, X_2) : P_t(X_1, V_1) \wedge P_t(X_2, V_2) \wedge belongs(v_1, V_1)$$

$$\wedge belongs(v_2, V_2) \wedge \neg belongs(v_1, V_2) \wedge \neg belongs(v_2, V_1)$$

où $belongs(v, V)$ est une fonction qui vérifie que v appartient à l'ensemble des valeurs V.

Traitement numérique. Soient X_1 et X_2 deux instances d'une classe, un chemin de propriétés P_t et deux ensembles de valeurs V_1 et V_2 du chemin P_t pour X_1 et X_2 respectivement. Un traitement numérique T_n est défini par :

$$T_n(X_1, X_2) : P_t(X_1, V_1) \wedge P_t(X_2, V_2) \wedge$$

$$compare_{Tn}(s(V_1), s(V_2))$$

où $compare_{Tn}$ est une fonction de comparaison de valeurs numériques pouvant être implémentée par $lessThan$ ou $greaterThan$ et s est une fonction d'agrégation qui peut être par exemple max, min, sum, etc.
Par exemple, un traitement T_c peut être que deux athlètes ont des manualités différentes : l'un est droitier et le second est gaucher. Un traitement T_n exprime une différence sur une valeur numérique, par exemple que le budget du club d'un athlète est plus élevé que celui d'un autre athlète.

Résultat. Nous considérons un chemin de propriétés P_o sous la forme $P_o : p_1(X, Z_1) \wedge p_2(Z_1, Z_2) \wedge \ldots \wedge p_m(Z_{m-1}, Z_m)$. Pour les résultats, nous considérons seulement les chemins de propriétés menant à des valeurs numériques. Soient X_1 et X_2 deux instances d'une classe, un chemin de propriétés P_o et deux ensembles de valeurs numériques V_1 et V_2 du chemin P_o pour X_1 et X_2 respectivement. Le résultat O est défini par :

$$O(X_1, X_2) : P_o(X_1, V_1) \wedge P_o(X_2, V_2) \wedge$$

$$lessThan(s(V_1), s(V_2))$$

où s est une fonction d'agrégation.

Règle Différentielle Causale. Une règle causale différentielle RDC_T représente la relation de causalité entre le traitement T et son résultat. Elle exprime que le traitement, i.e. une différence de valeurs sur un chemin de propriétés P_t, explique un résultat, i.e. une différence de valeurs sur un chemin de propriétés P_o tel que $lessThan(P_o(s(V_1)), P_o(s(V_2)))$.

Définition 1. (Règle Différentielle Causale). Étant données X_1 et X_2 deux instances d'une classe cible de l'ontologie, le chemin de propriétés menant au résultat P_o, un traitement $T \in \{T_n(X_1, X_2), T_c(X_1, X_2)\}$ défini par le chemin de propriété P_t, et s une fonction d'agrégation, une règle différentielle causale RDC_T est définie comme suit :

$$\begin{aligned} RDC_T : T \wedge P_o(X_1, V_1) \wedge P_o(X_2, V_2) \\ \Rightarrow lessThan(s(V_1), s(V_2)) \end{aligned} \tag{1}$$

Il est à noter que le résultat O est exprimé en partie dans le corps de la règle et dans sa conclusion. Une règle impliquant un traitement numérique est appelée une *règle différentielle causale numérique* et de manière analogue une règle impliquant un traitement catégoriel est appelée une *règle différentielle causale catégorielle*.

Exemple. Étant donné un graphe de connaissances décrivant les athlètes, leurs pays et leur sport, un résultat à étudier pourrait être le classement des athlètes, *i.e.* l'on va chercher à expliquer pourquoi des athlètes ont des performances différentes. Une règle différentielle causale numérique RDC_{age} peut exprimer qu'être plus jeune qu'un autre athlète peut expliquer un meilleur classement, *i.e.* un rang plus bas : $age(X_1, Y_1) \wedge age(X_2, Y_2) \wedge lessThan(Y_1, Y_2) \wedge rank(X_1, Z_1) \wedge rank(X_2, Z_2) \Rightarrow lessThan(Z_1, Z_2)$. Une règle différentielle causale catégorielle $RDC_{manualite}$ peut indiquer qu'être gaucher plutôt que droitier pourrait être une autre explication d'une meilleure performance.

3.2 Effet d'un Traitement

L'effet d'un traitement d'une règle est quantifié en calculant un score par la mesure $causal_T$ inspirée de [9]. Cette mesure correspond à un odds ratio OR qui compare les chances qu'un résultat se produise en fonction d'une exposition particulière, et les chances que le résultat se produise en l'absence de cette exposition. Alors que l'OR considère un ensemble d'instances, $causal_T$ considère seulement un ensemble de paires d'instances similaires et vérifiant le traitement.

Compte tenu des définitions des règles différentielles causales, nous définissons d'abord les deux supports utilisés pour calculer $causal_T$. Soit T un traitement tel que $T \in \{T_n(X_1, X_2), T_c(X_1, X_2)\}$, et O un résultat, $supp_{TO}$ représente le nombre de paires d'instances telles que le traitement et le résultat sont tous deux vérifiés :

$$supp_{TO} = \#(X_1, X_2) : \exists\{Y_1, .., Z_m\}\, tq\, T \wedge O(X_1, X_2) \tag{2}$$

Le support $supp_{T\overline{O}}$ représente le nombre de paires d'instances où le traitement et l'inverse du résultat sont vérifiés, *i.e.* avec le chemin de propriété P_o instancié pour les deux instances mais des valeurs numériques qui vérifient le prédicat $greaterThan$ au lieu de $lessThan$. Il est calculé de manière analogue à celui de $supp_{TO}$, sauf que le résultat O est remplacé par $\overline{O}$, avec $\overline{O} \equiv P_o(X_1, V_1) \wedge P_o(X_2, V_2) \wedge greaterThan(s(V_1), s(V_2))$.
Étant donné un ensemble de paires, la mesure de l'effet d'un traitement est défini dans l'équation 3 suivante :

$$causal_T = \frac{supp_{TO}}{supp_{T\overline{O}}} \tag{3}$$

$causal_T$ retourne un score dans $[0, +\infty[$ et mesure la force de la relation entre le traitement et le résultat. S'il est égal à 1, le traitement et le résultat sont considérés comme indépendants, et plus il est différent de 1, plus cette relation est forte. Une valeur supérieure à 1 montre que le traitement et le résultat sont positivement associés.
Nous utilisons cette mesure pour sélectionner les règles pertinentes et ordonner les explications fournies. Un intervalle de confiance CI est construit pour tester si $causal_T$ est significativement supérieur à 1 : $CI_\alpha(causal_T) = exp(ln(causal_T) \pm u_{1-\alpha/2}\sqrt{\frac{1}{supp_{TO}} + \frac{1}{supp_{T\overline{O}}}})$ avec $u_{1-\alpha/2}$ le $(1-\alpha/2)$ quantile de la loi normale $\mathcal{N}(0, 1)$.

4 Appariement d'instances fondé sur les plongements de graphes

L'effet d'une règle est calculé en considérant des paires d'instances similaires. Comme il a été mentionné dans la section 2, les méthodes classiques d'appariement symbolique peuvent échouer lorsque les descriptions des instances sont hétérogènes, erronées ou incomplètes. Pour notre problème, nous avons besoin d'une mesure de similarité capable de déterminer des instances similaires même lorsque le graphe de connaissances est imparfait. Pour ce faire, nous avons défini une nouvelle mesure de similarité qui exploite les plongements de graphes dans un espace vectoriel à faible dimensions. Plus précisément, nous entraînons un modèle de plongements pour fournir un vecteur à chaque instance. Ensuite, pour mesurer la similarité de deux vecteurs, nous analysons la similarité des prédictions obtenues en utilisant ces vecteurs.
Dans la figure 1, nous présentons le déroulement général de notre approche en cinq étapes. Tout d'abord, (a) nous considérons un graphe de connaissances KG et entraînons un modèle de plongements de graphe qui fournit une représentation vectorielle de chaque instance et relation du KG dans un espace de faible dimension. Ensuite, dans (b), un ensemble de paires est sélectionné aléatoirement, puis la distance de leur vecteurs d et leur similarité sim_e (c.f. équation 4 en section 4.1) sont calculées. Dans (c), un modèle g, en considérant d et sim_e est appris pour définir le seuil de distance d_{tr} étant donné un paramètre sim_{tr}. À (d), un ensemble d'instances appariées, *i.e.* avec une distance d telle que $d < d_{tr}$, est alors créé. Enfin, à (e), nous calculons l'effet de traitement sur le résultat analysé.

4.1 Mesures de similarité sur plongements

Distance basée sur des vecteurs de plongements. Étant donné un KG, nous entraînons un modèle de plongements pour obtenir des vecteurs qui représentent les entités et les relations du KG dans un espace de faible dimension.
Une distance d entre deux vecteurs v_{i_1} et v_{i_2} peut être calculée à l'aide de diverses fonctions telles que la distance euclidienne. Une telle distance est une valeur dans $[0, +\infty[$. Cette valeur peut être utilisée pour décider que deux instances sont similaires en utilisant un seuil donné d_{tr} (*i.e.* i_1 et i_2 considérés similaires si $d(v_{i_1}, v_{i_2}) < d_{tr}$).
Cependant, la définition d'un tel seuil d_{tr} est difficile sans connaissance de ce que les distances représentent. Sans seuil défini, l'étape d'appariement pourrait sélectionner, pour une instance i_1, i_2 tel que $\min_{\forall i_x \in I} d(v_{i_1}, v_{i_x})$, mais i_1 et i_2 pourraient être très différents. Par exemple, un KG pourrait comporter des descriptions de pays comme la France, les États-Unis et l'ensemble des pays d'Asie. En supposant que l'on cherche un pays similaire à la France, l'on pourrait trouver les États-Unis. Cependant, ces 2 pays restent très différents, et ne pourraient peut-être pas être considérés similaires pour notre problème.
Ainsi, un seuil d_{tr} est à définir pour (i) sélectionner des paires d'instances suffisamment similaires pour analyser un traitement, et (ii) élaguer les paires trop différentes.

Mesures de similarité basées sur des prédictions de plongements. Nous proposons une nouvelle mesure de similarité, sim_e, qui repose sur les plongements des instances. sim_e mesure la similarité de deux vecteurs appris en observant la similarité des prédictions utilisant ces vecteurs. Nous supposons que deux instances ayant deux descriptions RDF similaires conduiront à des vecteurs similaires dans l'espace de plongement [22], et que ces vecteurs produiront des prédictions similaires de la part du modèle de plongement. Nous rappelons qu'un modèle de plongement f reçoit en entrée un triplet (h, r, t), et retourne un score $f_r(h, t)$ qui sera élevé si le modèle considère le triplet correct, et bas avec un triplet considéré faux. La principale motivation de cette mesure est que la comparaison des vecteurs permettra à l'approche d'être moins sensible aux données incomplètes ou erronées.
Plus précisément, étant donné un sujet et une propriété p_k, un modèle de plongements f prédit un score pour chaque objet possible de p_k et peut trier les objets par score décroissant. Ainsi, deux instances similaires devraient avoir des prédictions similaires sur les propriétés qu'elles instancient.

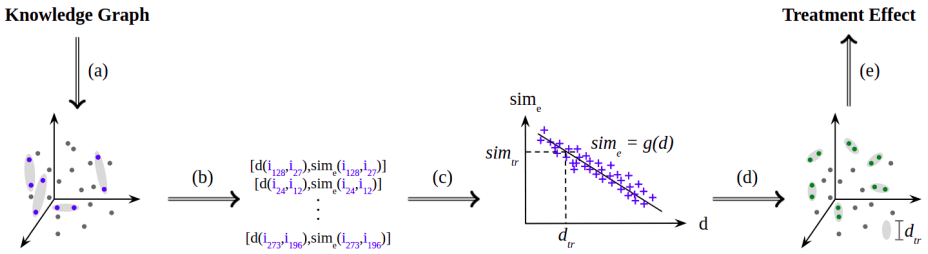


FIGURE 1 – Vision globale de l'approche

Pour calculer sim_e, les prédictions sur chaque propriété p_k instanciant les instances sont étudiées. Pour une propriété p_k d'un chemin P de longueur $l_P < l_{max}$, un degré de fonctionnalité $degre(p_k)$ est défini et représente le nombre moyen d'objets distincts qui instancient p_k. l_{max} est un paramètre qui définit la longueur maximale d'étude d'un chemin de propriétés. Avec $n = degre(p_k)$, la similarité entre deux instances sur la propriété p_k est étudiée en analysant les n premiers objets prédits par f pour chaque instance. Par exemple, pour les pays visités par une personne, nous pouvons poser le degré de fonctionnalité $n = 3$. Ainsi, les 3 pays les mieux classés pour chaque personne sont sélectionnés et utilisés. La similarité sim_{p_k} de deux instances i et j sur une propriété p_k de la classe cible est calculée récursivement comme suit :

$$sim_{p_k}(i,j) = \begin{cases} 1, \ si \ p_k(i) = p_k(j) \\ 0, \ si \ p_k(i) \neq p_k(j), \ i \in L, \ j \in L \\ \\ \frac{\sum_{o_i \in p_k(i)} Max_{o_j \in p_k(j)} sim_e(o_i, o_j)}{degre(p_k)} \\ si \ p_k(i) \neq p_k(j), \ i \in I, \ j \in I \end{cases}$$

où L (resp. I) est l'ensemble des littéraux (resp. des IRIs), $p_k(i)$ est l'ensemble des objets liés à i par p_k.
La similarité entre i et j, $sim_e(i,j)$, est obtenue en faisant la moyenne de la similarité obtenue pour chaque propriété p_k qui appartient à l'ensemble P des propriétés instanciées pour i et j :

$$sim_e(i,j) = \frac{\sum_{p_k \in P} sim_{p_k}(i,j)}{\mid P \mid} \tag{4}$$

Il convient de noter qu'un poids relatif w_i représentant l'importance d'une propriété p_i dans le calcul de similarité peut être introduit et défini par des experts du domaine (par exemple, w_i peut être fixé à 0 pour le nom d'une personne). Nous proposons un exemple pour illustrer sim_e basé sur la Fig. 2. Premièrement, $sim_{registered}(\#person1, \#person2) = \frac{sim_e(\#master1, \#master2)}{1} = \frac{sim_{hasTopic}(\#master1, \#master2)}{2} = \frac{(max(sim_e(IT,IT), sim_e(IT,Physics)) + max(sim_e(Maths,IT), sim_e(Maths,Physics)))}{2} = \frac{1}{2}$.
Ensuite, $sim_{citizenship}(\#person1, \#person2) = \frac{sim_e(\#greece1, \#greece1)}{1} = 1$. Enfin, $sim_e(\#person1, \#person2) = \frac{1/2}{2} + \frac{1}{2} = \frac{3}{4}$.

TABLE 1 – Distance entre vecteurs de plongement de 8 instances

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
i_1	*0,0*	2,7	1,6	1,9	**0,5**	2,1	2,9	2,7
i_2	2,7	*0,0*	1,6	2,0	3,1	**0,3**	2,5	1,4
i_3	1,6	1,6	*0,0*	1,4	2,2	2,4	3,2	2,3
i_4	1,9	2,0	1,4	*0,0*	2,5	1,8	2,4	2,5
i_5	**0,5**	3,1	2,2	2,5	*0,0*	1,3	3,6	0,9
i_6	2,1	**0,3**	2,4	1,8	1,3	*0,0*	0,8	2,9
i_7	2,9	2,5	3,2	2,4	3,6	0,8	*0,0*	**0,5**
i_8	2,7	1,4	2,3	2,5	0,9	2,9	**0,5**	*0,0*

4.2 Définition du seuil de similarité d_{tr}

d_{tr} est défini en analysant la distribution entre sim_e et d (parties (b) et (c) de la figure 1). Un modèle g entre d et sim_e, tel que $g(d) = sim_e$, est appris. Etant donné qu'aucune hypothèse n'est avancée sur la relation entre d et sim_e, g peut être un modèle linéaire ou non linéaire. Étant donné sim_{tr} le seuil défini sur la similarité sim_e fixé par l'utilisateur, le seuil d_{tr} est défini tel que $g(d_{tr}) = sim_{tr}$.
Le calcul de sim_e étant complexe en temps, la distribution entre sim_e et d est étudiée sur un échantillon de paires.
Dans la table 1, la distance entre 8 vecteurs de plongements d'instances d'une classe cible est affichée. Nous supposons que sim_{tr} a été fixé à 0,8, que le modèle $sim_e = -0,13 * d + 1$ a été appris et que d_{tr} est fixé à 1,3. Ainsi, 3 paires seraient sélectionnées.

5 Algorithme

L'algorithme Dicare-E est composé de 2 parties : (i) la première consiste à déterminer d_{tr}, et (ii) la seconde crée des appariements d'instances similaires en utilisant d_{tr} afin d'évaluer la règle différentielle causale pour un traitement.

5.1 Définition de d_{tr}

Dans la première partie de l'algorithme, d_{tr} est déterminé en analysant la relation entre la distance d et la métrique de similarité sim_e. À cette fin, l'algorithme prend en entrée le KG, la classe cible t_c, les chemins de propriétés liés au résultat et au traitement étudiés P_o et P_t, un modèle de plongement f et le paramètre sim_{tr} pour définir d_{tr}. Le graphe de connaissances KG_{tr}, utilisé pour entraîner f, est obtenu en retirant les derniers prédicats de P_o et P_t de KG. d_{tr} est défini de la façon suivante. Étant donné

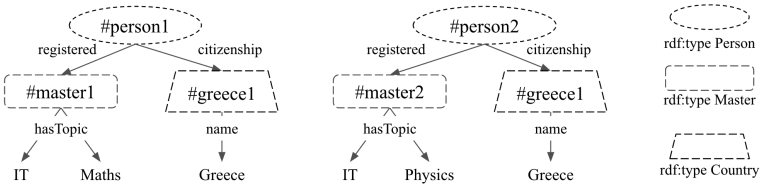


FIGURE 2 – Descriptions prédites de 2 instances pour illustrer sim_e

$\mathcal{I}_{tc}$ l'ensemble des instances de t_c, un échantillon de paires de t_c, $\{(i_i, i_j) \in \mathcal{I}_{tc} \times \mathcal{I}_{tc}\} \subset \mathcal{I}_{tc} \times \mathcal{I}_{tc}$, est tiré, et les distances d entre les vecteurs correspondants $d(v_{i_i}, v_{i_j})$ et les sim_e sont calculées pour chaque paire grâce au modèle appris f. Un second modèle, $g(d) = sim_e$, est entraîné pour obtenir les paramètres de la distribution entre d et sim_e. g peut être une régression linéaire ou polynomiale en fonction de la distribution, pouvant être linéaire ou non. Les paramètres estimés du modèle sont utilisés pour obtenir d_{tr} tel que $g(d_{tr}) = sim_{tr}$.

5.2 Création des paires et découverte de règles RDC_T

La deuxième partie de l'algorithme consiste à déterminer des règles causales différentielles en utilisant d_{tr} précédemment défini. En entrée, l'algorithme considère la classe cible t_c, les chemins P_o et P_t, d_{tr} et α qui est un paramètre statistique utilisé pour calculer un intervalle de confiance pour les effets du traitement.

L'extraction des règles se fait en deux étapes. La première étape consiste à construire les paires d'instances similaires vérifiant T grâce à d_{tr}. Premièrement, pour chaque paire $(i_i, i_j) \in \mathcal{I}_{tc} \times \mathcal{I}_{tc}$, la distance euclidienne entre leurs vecteurs $d(v_{i_i}, v_{i_j})$ est calculée. Les distances sont stockées dans une matrice de distance $\mathcal{D}$. Ensuite, l'ensemble d'instances appariées $\mathcal{M}$ est créé en sélectionnant à chaque itération la paire présentant la plus petite distance et vérifiant le traitement, $\min_{\forall (i_i, i_j) \in \mathcal{I}_{tc} \times \mathcal{I}_{tc}} \mathcal{D}$ et en l'enlevant de $\mathcal{D}$ ensuite. Ce processus est appliqué jusqu'à ce qu'il ne reste aucune paire (i_i, i_j) dans $\mathcal{D}$ telle que $d(v_{i_i}, v_{i_j}) < d_{tr}$.

La deuxième étape consiste à calculer l'effet du traitement. Pour cela, l'algorithme prend en entrée l'ensemble des paires d'instances $\mathcal{M}$, les chemins P_o et P_t, et α, et est initialisé en fixant à 0 les deux supports décrits dans l'équation 3. Ensuite, pour chaque paire, le traitement et le résultat sont obtenus en utilisant P_o et P_t, et les supports sont modifiés en conséquence. Une fois que toutes les paires ont été traitées, l'effet du traitement peut être calculé et son intervalle de confiance construit.

Cet algorithme permet de déterminer l'effet d'un traitement. Afin de déterminer l'effet d'un autre traitement, KG_{tr} doit être mis à jour en enlevant le nouveau traitement et en ajoutant l'ancien, et f réentrainé en conséquence.

TABLE 2 – Description des Données

	DBPediaW	***Vitamin***
# Triplets	6908	86006
# Classes	4	19
# Instances t_c	185	1714
# Propriétés	8	22

6 Expériences

6.1 Données Utilisées

Il n'existe pas de référence pour la découverte de causalité dans les KGs. Nous avons exploité deux KGs déjà utilisés : *Vitamin* [20], pour lequel un expert du domaine est disponible pour une évaluation qualitative, et un extrait relativement simple de *DBPedia*, que l'on nomme *DBPediaW*, utilisé dans [11] et [20]. Les informations de ces graphes sont présentés dans la table 2.
Vitamin décrit des personnes et leurs caractéristiques socio-économiques telles que leur âge, leur lieu de vie, leur travail, leur sexe, leur régime alimentaire actuel et idéal, leurs opinions sur des faits liés au bien-être animal et au changement climatique. Ce graphe a une profondeur de 2. La classe cible $Person$ a 1714 instances, et nous souhaitons expliquer la différence entre le régime actuel et idéal d'une personne, *i.e.* sa volonté de réduire sa consommation de viande, indiquée par le prédicat *reduceMeat* ayant des valeurs $\in \mathbb{N}$ Nous nous concentrons sur la recherche de traitements qui pourraient expliquer la volonté d'une personne de changer ses habitudes alimentaires.
Le lecteur est invité à consulter [11] pour visualiser le schéma de *DBPediaW*. Ce graphe a une profondeur de 2 et décrit des auteurs, leur parcours académique et des livres qu'ils ont écrits. La classe $Writer$ est composée de 185 instances. Nous cherchons à expliquer pourquoi certains auteurs publient leur premier livre plus jeunes que d'autres.

6.2 Évaluation et Résultats

L'objectif est de montrer l'efficacité et la robustesse de l'utilisation de modèles de plongements de graphes pour la découverte de règles causales différentielles dans les KGs.
Tout d'abord, nous avons entraîné et évalué les modèles suivants sur les KGs : *TransE*, *DistMult*, *ComplEx*, *HolE*, *ConvE* et *ConvKB*. Ensuite, nous avons étudié la distribution entre la distance euclidienne d et la métrique de similarité sim_e afin de montrer que cette distribution peut être

TABLE 3 – Résultats

	DBPediaW	***Vitamin***
# Règles ([19])	12	0
# Règles ([20])	12	77
# Règles Dicare-E	3	48
% Paires expliquées ([19])	21,2	0
% Paires expliquées ([20])	21,2	50.1
% Paires expliquées Dicare-E	**78,1**	**92,8**
% Règles pertinentes ([20])	*NA*	66,6
% Règles pertinentes Dicare-E	*NA*	**76,6**

TABLE 4 – Performance des modèles entrainés sur *Vitamin*

Model	***MRR***	***Hits@1***	***Hits@3***	***Hits@10***
ConvE	**0,3384**	**0,2498**	**0,3969**	**0,4850**
DistMult	0,2167	0,1341	0,2451	0,3675
ComplEx	0,1732	0,1011	0,1867	0,3113
TransE	0,1470	0,1341	0,1563	0,2474
Baseline	*0,0057*	*0,0009*	*0,0023*	*0,0081*

estimée par un modèle. Enfin, nous avons appliqué notre algorithme pour découvrir un ensemble de règles causales différentielles qui représentent des effets de traitements, et comparons les résultats obtenus aux règles différentielles causales obtenues par [19] - appariement strict - et [20] - appariement par communautés. Plus précisément, l'objectif est de comparer l'interprétabilité et la pertinence des règles obtenues, le nombre de paires d'instances qu'elles peuvent expliquer, et la robustesse des deux approches en cas de données incomplètes.

Entraînement des modèles de plongements. Le tableau 4 montre que, parmi tous les modèles testés à l'aide de la librairie AmpliGraph sur *Vitamin*, $ConvE$ est le plus performant, *i.e.* avec les *MRR* et *Hits@n* les plus élevés, et est donc utilisé par la suite. *DistMult* est utilisé pour représenter les vecteurs de *DBPediaW* car il obtient les meilleurs performances. Par soucis de représentation des résultats, l'ensemble des tables n'est pas présenté mais peut être trouvé à ce lien [1].

Association entre distance d et similarité sim_e. Pour chaque KG, un ensemble de paires d'instances est tiré aléatoirement et les valeurs d et sim_e sont calculées pour chaque paire. La distribution entre d et sim_e a été modélisée par un modèle linéaire dans les 2 cas. Pour *Vitamin*, la distribution entre d et sim_e est présentée dans la Fig. 3 (à gauche). Le même processus est effectué avec la baseline (à droite). Comme prévu par [22], plus la distance d entre les vecteurs de deux instances est élevée, moins les instances sont similaires, car sim_e diminue lorsque d augmente. La comparaison avec la baseline montre que $ConvE$ est capable de capturer correctement les représentations des entités de *Vitamin* en rapprochant les entités similaires dans l'espace de plongements, et que l'utilisation d'un modèle avec de bonnes performances est nécessaire pour cette approche. La distribution entre d et sim_e de la Fig. 3 est modélisée par un modèle linéaire avec $(r^2 = 0,78)$. En utilisant ce modèle, le seuil de distance d_{tr} est fixé à $1,1$ pour obtenir une similarité sim_{tr} de 0,75.

Règles découvertes et effet de d_{tr}. Deux traitements différents sont analysés pour illustrer l'importance de la détermination de d_{tr} dans la Fig. 4. : le sexe et le lieu de vie de personnes. Une règle est considérée valable si la barre d'erreur ne croise pas la barre horizontale placée à 1. Cette figure montre que plus d_{tr} est faible, plus les barres d'erreurs sont larges. Les paires d'instances des ensembles obtenus sont très similaires, mais la taille de ces ensembles diminuent avec d_{tr}, résultant en une variance élevée dans l'effet estimé. Inversement, plus d_{tr} est élevé, plus les ensembles sont grands mais avec des paires moins similaires. En conséquence, cela mène à une plus faible variance de l'effet et à un biais plus élevé [21]. Le fait que d_{tr} soit fixé avant le calcul des effets de traitement évite d'introduire un biais dans les règles de la part de l'utilisateur. La Fig. 4 nous indique que, pour d_{tr} fixé à 1.1, la volonté de réduire sa consommation de viande est indépendante du genre car la barre d'erreur correspondante croise la barre horizontale. Il semble en revanche y avoir un effet du lieu d'habitation, les barres ne se croisant pas, habiter en campagne plutôt qu'en aire urbaine pourrait expliquer une volonté de réduire sa consommation de viande.

$$\begin{aligned} RDC_{hasDiet} : hasDiet(X_1, omnivorous) \wedge \\ hasDiet(X_2, vegetarian) \wedge reducing(X_1, Z_1) \wedge \\ reducing(X_2, Z_2) \Rightarrow lessThan(Z_2, Z_1) \end{aligned} \quad (5)$$

$$\begin{aligned} RDC_{bornIn} : bornIn(X_1, Y_1) \wedge bornIn(X_2, Y_2) \wedge \\ publishedIn(X_1, Z_1) \wedge publishedIn(X_2, Z_2) \\ \wedge greaterThan(Y_2, Y_1) \Rightarrow lessThan(Z_2, Z_1) \end{aligned} \quad (6)$$

Évaluation qualitative et quantitative et comparaison. Pour évaluer notre approche et la comparer à l'état de l'art [19, 20], trois critères ont été évalués. Qualitativement, les règles des deux approches ont été évaluées. Ensuite, nous avons étudié le pourcentage de paires pour lesquelles les approches peuvent fournir une explication concernant une différence de régimes. Enfin, nous avons testé la robustesse des deux approches aux données incomplètes.

Notre approche découvre 48 règles différentielles causales pour *Vitamin*, et 3 pour *DBPediaW*. Nous présentons 2 règles dans les équations 5 et 6 [2] et invitons le lecteur à visiter le GitHub pour obtenir la liste exhaustive des règles. La règle de l'équation 5 exprime qu'être omnivore par rapport à être végétarien explique une volonté plus forte de réduire sa consommation de viande. Cette règle fait sens car, pour A et B deux personnes similaires, si A consomme déjà moins de produits animaux que B, alors B est plus susceptible de réduire sa consommation de produits animaux.

1. https ://github.com/IC2022RuleEmbeddings/Soumission

2. La représentation des chemins a été simplifiée, les *belong* sont omis, et la fonction s n'est pas définie car les propriétés numériques sont monovaluées dans $Vitamin$ et $DBPediaW$.

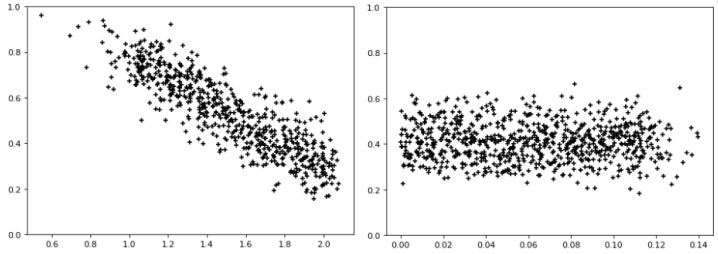


FIGURE 3 – Distance (axe x) et similarité (axe y) d'un ensemble de paires pour $ConvE$ (gauche) et $Baseline$ (droite)

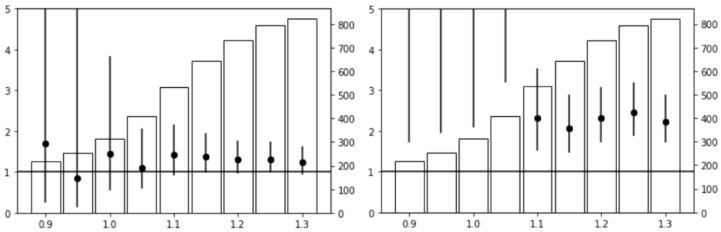


FIGURE 4 – $causal_{OR_T}$ (axe y gauche, points) et nombre de paires créées (axe y droite, histogramme) selon d_{tr} (axe x). Traitements : (gauche) genre (X_1 : femme, X_2 : homme) ; (droite) lieu d'habitation (X_1 : campagne, X_2 : aire urbaine)

La règle de l'équation 6 exprime qu'être né plus tard qu'un autre auteur explique une publication à un âge plus jeune de son premier livre. Ceci peut être expliqué par une accessibilité grandissante de la publication due à un plus grand nombre d'éditeur ou encore aux réductions de coûts de publication.

Vitamin. Les règles découvertes par [20] et Dicare-E ont été évaluées par un expert du domaine, en comportement humain et nutrition. Aucune instance n'a pu être appariée avec [19] qui nécessite que leurs représentations soit isomorphes au traitement et au résultat près. En conséquence, aucune règle n'a pu être extraite (voir table 6.1). Pour faciliter le processus de comparaison des approches, les 30 meilleures règles de [20] et Dicare-E ont été évaluées.

Pour rappel, les règles générées dans [20] peuvent utiliser un motif de graphe plus ou moins spécifique, nommé strate, qui définit l'ensemble des instances auxquelles la règle est appliquée. Dans les deux approches, toutes les règles sont interprétables par l'expert. L'expert a souligné que plus la strate est spécifique, plus elles sont difficiles à interpréter, alors que la plupart des règles de Dicare-E sont faciles à interpréter.

Pour chaque règle, quatre possibilités ont été données à l'expert : la règle (1) *semble être pertinente*, (2) *pourrait être pertinente*, (3) l'expert *ne sait pas si la règle est pertinente* et (4) *semble être fausse*. Cependant, elles diffèrent dans la distribution : pour Dicare-E, 11 règles ont été classées dans (1), 12 dans (2), 6 dans (3) et 1 dans (4). Pour les règles de [20], la distribution est la suivante : 12 règles en (1), 8 en (2), 6 en (3) et 4 en (4). Par conséquent, 66,6% des règles de [20] semblent être pertinentes ou pourraient l'être, ce qui est inférieur à notre approche (76,6%). De plus, Dicare-E n'a qu'une seule règle qui semble fausse, alors que [20] en a 4. La justification de l'expert sur la règle semblant être fausse est que les valeurs de traitement sont trop similaires pour que la règle ait un sens : les deux font référence à l'importance de l'entourage d'une personne dans sa prise de décision, l'une reposant davantage que l'autre sur son entourage. Par ailleurs, notre approche fournit un ensemble plus large de traitements : des chemins de propriété supplémentaires par rapport à [20] ont été déterminés.

Pour une évaluation quantitative, le pourcentage de paires pour lesquelles au moins une explication du résultat peut être fournie a été calculé pour les deux approches. Notre approche peut fournir une explication pour 92,8% de ces paires alors que [20] obtient un nombre de 50,1%, elle explique donc beaucoup plus de différences. Nous avons calculé la même métrique en supprimant 15% des triplets de *Vitamin* pour tester la robustesse des deux approches. Le nombre de règles extraites et le pourcentage de paires expliquées ont été réduits dans les deux approches, mais notre approche est plus robuste : le pourcentage de paires expliquées est passé de 92,8% à 89,5% pour notre approche, et de 50,1% à 24,7% pour [20]. Ceci s'explique du fait qu'une instance avec une description incomplète ne sera pas utilisée pour la détermination de règles dans [20] alors qu'elle le sera dans notre approche.

DBPediaW. Les 3 règles obtenues par Dicare-E sont compréhensibles et semblent pertinentes. De plus, elles comportent des traitements également obtenus dans [20]. Sur les 12 règles obtenues par [20], 9 semblent pertinentes. Les résultats de [19] sont les mêmes que [20] car, le schéma étant simple, aucune détection de communauté n'a été réalisée. Il est intéressant de voir que par opposition au jeu de données précédent, pour un KG comme *DBPediaW* dont le schéma est relativement simple, les règles déterminées par [20] ont des strates avec peu d'éléments qui sont donc facilement compréhensibles et qui permettent de générer des

règles plus expressives que Dicare-E (*e.g.* règle concernant la date de naissance valide pour les auteurs ayant étudié aux États-Unis dans une université bien classée).
Quantitativement, notre approche permet d'expliquer 78,1% des paires contre 21,2% pour [20]. En enlevant 15% des triplets, les règles obtenues avec notre approche sont les mêmes. Le nombre de paires expliquées ne change donc pas. Avec [20], ce nombre passe à 6,0%. Comme sur *Vitamin*, notre approche permet d'expliquer plus de paires ayant une différence de résultat et est plus robuste aux données manquantes.

7 Conclusion

Dans cet article, nous avons proposé une approche qui combine des modèles de plongements de graphes et des techniques de fouille de règles symboliques pour la découverte de règles différentielles causales dans les graphes de connaissances. Une telle approche hybride est capable de traiter efficacement des données incomplètes tout en fournissant des règles interprétables, qui peuvent expliquer les différences dans une caractéristique numérique étudiée. Nous avons montré qu'elle peut être utilisée pour apparier des instances similaires grâce à leur représentation dans l'espace des plongements appris, permettant ainsi l'application du cadre des résultats potentiels. La métrique de similarité proposée, basée sur les prédictions du modèle de plongements, garantit la création de paires similaires uniquement. Notre expérience et collaboration avec un expert montre que notre approche peut être utilisée sur des domaines variés ainsi que sur des KGs complexes. Dans de futurs travaux, nous souhaitons analyser les règles plus en profondeur, notamment le nombre de règles pouvant expliquer le résultat d'une paire et ses conséquences.

Références

[1] Althauser, R.P., Rubin, D. : The Computerized Construction of a Matched Sample. American Journal of Sociology 76(2), 325–346 (1970)

[2] Chickering, D.M., Heckerman, D., Meek, C. : Large-sample learning of Bayesian networks is NP-hard. JMLR 5, 1287–1330 (2004)

[3] Ferilli, S., Basile, T.M., Biba, M., Di Mauro, N., Esposito, F. : A general similarity framework for horn clause logic. Fundamenta Informaticae 90(1-2), 43–66 (2009)

[4] Ferré, S. : Answers Partitioning and Lazy Joins for Efficient Query Relaxation and Application to Similarity Search. Lecture Notes in Computer Science 10843 LNCS, 209–224 (2018)

[5] Galárraga, Luis Teflioudi, Christina Hose, Katja Suchanek, Fabian. (2013). AMIE : Association rule mining under incomplete evidence in ontological knowledge bases. WWW 2013. 413-422.

[6] Haan, Rosaline Tiddi, Ilaria Beek, Wouter. Discovering Research Hypotheses in Social Science Using Knowledge Graph Embeddings. The Semantic Web 477-494 (2021)

[7] Iacus, S.M., King, G., Porro, G. : Causal inference without balance checking : Coarsened exact matching. Political Analysis 20(1), 1–24 (2012)

[8] Jain, N., Kalo, J.C., Balke, W.T., Krestel, R. : Do embeddings actually capture knowledge graph semantics ? In : Verborgh, R., Hose, K., Paulheim, H., Champin, P.A., Maleshkova, M., Corcho, O., Ristoski, P., Alam, M. (eds.) The Semantic Web. pp. 143–159.

[9] Li, Jiuyong le, Thuc Liu, Lin Liu, Jixue Jin, Zhou Sun, Bingyu. (2013). Mining Causal Association Rules. ICDMW 2013. 114-123.

[10] Moon, C., Jones, P., Samatova, N.F. : Learning entity type embeddings for knowledge graph completion. CIKM '17 p. 2215–2218. ACM, NY, USA (2017)

[11] Munch, M., Dibie, J., Wuillemin, P., Manfredotti, C.E. : Towards interactive causal relation discovery driven by an ontology. In : International Florida Artificial Intelligence Research Society Conference (2019)

[12] Neapolitan, R.E. : Learning Bayesian Networks. Pearson Prentice Hall. (2003)

[13] Nentwig, M., Hartung, M., Ngomo, A.N., Rahm, E. : A survey of current link discovery frameworks. Semantic Web 8(3), 419–436 (2017)

[14] Paulheim, H. : Knowledge graph refinement : A survey of approaches and evaluation methods. Semantic Web 8(3), 489–508 (2017).

[15] Pearl, J. : Causality. Cambridge University Press (2009)

[16] Raad J., Pernelle N., Saïs F. Detection of Contextual Identity Links in a Knowledge Base. In Proceedings of the Knowledge Capture Conference (K-CAP 2017). Association for Computing Machinery

[17] Rosenbaum, P.R., Rubin, D.B. : Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 79(387), 516–524 (1984)

[18] Rubin D. B : Estimating causal effects of treatment in randomized and nonrandomized studies. Journal of Educational Psychology 66(5), 688–701 (1974)

[19] Simonne, L., Pernelle, N., Sais, F. : Fouille de règles différentielles causales dans les graphes de connaissances, EGC 2021, pp.293-300

[20] Simonne L., Pernelle N, Saïs F., Thomopoulos R. Differential Causal Rules Mining in Knowledge Graphs. In Proceedings of the 11th on Knowledge Capture Conference (K-CAP '21). Association for Computing Machinery, 105–112.

[21] Stuart, E.A. : Matching methods for causal inference : A review and a look forward. Statistical science : a review journal of the Institute of Mathematical Statistics 25(1), 1–21 (2010)

[22] Wang, C., Pan, S., Hu, R., Long, G., Jiang, J., Zhang, C. : Attributed graph clustering : A deep attentional embedding approach. IJCAI 2019 (2019), 3670-3676.

Identifier et reconnaitre des essences de bois à l'aide d'un réseau Bayésien basé sur des indicateurs macroscopiques

C. Baudrit[1], F. Michaud[2], C. Fernandez[1], M. Munch[1]

[1] INRAE, Université de Bordeaux-I2M, Talence, France
[2] LIMBHA, Ecole Supérieure du Bois, Nantes, France

cedric.baudrit@inrae.fr, franck.michaud@esb-campus.fr, christophe.fernandez@inrae.fr, melanie.munch@u-bordeaux.fr

Résumé

La reconnaissance et/ou l'identification d'une essence de bois, à l'échelle d'une pièce, nécessite de pouvoir manipuler, agréger et intégrer un certain nombre de variables hétérogènes simultanément. Les outils développés s'appuient principalement sur de l'analyse d'image ou des mesures physiques nécessitant des équipements spécifiques difficilement utilisables dans la pratique ou qui agissent comme une boite noire pour l'utilisateur. Un modèle, basé sur les réseaux Bayésiens, est proposé pour (1) guider, en situation de pratique professionnelle sur site, un utilisateur, un apprenant ou un expert dans l'identification d'une essence au travers d'un ensemble de questions et/ou (2) fournir des conclusions et des propositions pertinentes de descripteurs, caractéristiques les plus vraisemblables d'une essence (re)connue.

Mots-clés

Incertitude, Réseaux Bayésiens, Transfert et ingénierie de connaissances, Essence de bois.

Abstract

The identification of wood species from samples has mainly relied on image analysis. The tools developed using this approach are often difficult to use and require specific equipment in order to acquire the specific images and the range of data needed. To overcome these drawbacks, we have developed a versatile and practical tool, based on the formalism of Bayesian networks. It allows users to both identify and recognize wood species in a sample while also serving as a knowledge base for the transfer of knowledge in a learning context. Using a set of macroscopic descriptors, the tool guides the user in the identification of the wood species and from an identified wood species, it provides a map of its characteristics. The tool being both predictive and instructive, is useful for professionals in the wood industry as well as for training purposes in educational facilities.

Keywords

Uncertainty, Bayesian networks, knowledge engineering and transfer, wood species identification.

1 Introduction

La reconnaissance des essences de bois dans un morceau de bois ou un objet en bois et l'identification des caractéristiques anatomiques de ces essences sont des problèmes difficiles dans la mesure où le bois est un matériau complexe et hétérogène. D'ailleurs il ne s'agit pas seulement d'un matériau unique mais bien d'une grande diversité d'essences (plusieurs dizaines de milliers) qui présente une grande variabilité qu'elle soit inter ou intraspécifique, qu'elle soit inter ou intra arbres. La reconnaissance et/ou l'identification d'une essence de bois, à l'échelle d'un échantillon ou d'une pièce, requiert la capacité de gérer, d'agréger et d'intégrer simultanément des variables hétérogènes [5]. Actuellement, l'identification des essences de bois repose presque exclusivement sur l'expertise et est réalisée par des spécialistes expérimentés à l'aide de caractéristiques macroscopiques et microscopiques telles que la couleur, la structure et la texture. Afin (1) de réduire le coût et le temps nécessaire à la formation des experts, (2) de surmonter les limites du mécanisme cognitif utilisé par les experts pour agréger les caractéristiques et (3) d'améliorer la précision de la reconnaissance des essences de bois, des approches automatiques ou semi-automatiques ont été proposées. La grande majorité de ces approches sont basées sur des méthodes d'analyses d'images [1, 7, 10, 22, 24]. Cependant, la pièce de bois observée a souvent été modifiée par rapport à son origine, ou présente un état de surface qui rend l'analyse d'image difficile à utiliser. Des outils, basés sur la génétique du bois ou des mesures physiques, ont également été mis en place [11, 19]. Ces outils nécessitent alors des équipements spécifiques pour l'acquisition d'images ou de données à différentes échelles qui sont difficiles à mettre en place dans la réalité. Certaine récente avancée combine, dans un outil dédié au terrain, les technologies de façon efficace sans toutefois considérer la notion d'apprentissage de l'utilisateur [18]. De plus, toutes ces méthodes reposent souvent sur des modèles de type "boîte noire" difficiles à interpréter. Il est donc nécessaire de proposer un outil capable d'assembler des connaissances hétérogènes dans le but de guider un utilisateur, un apprenant ou un expert sur site dans l'identification des essences de bois tout en fournissant une

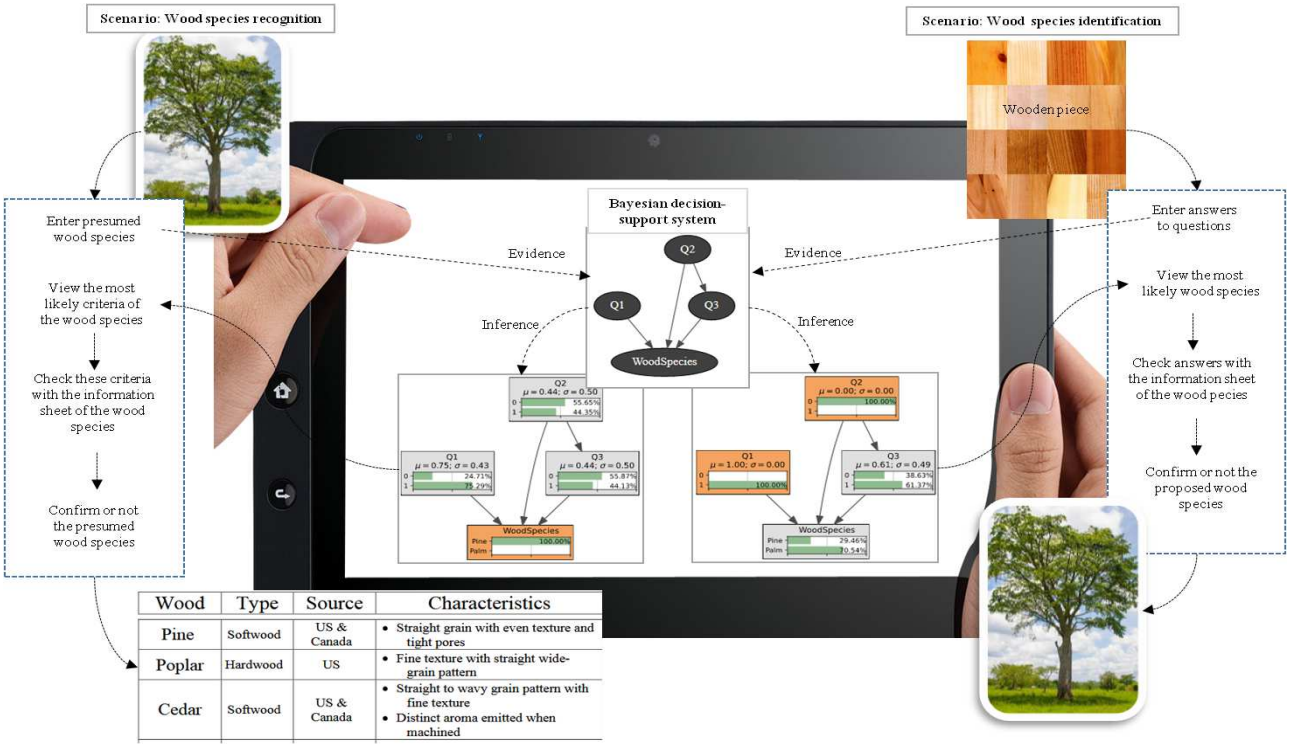


FIGURE 1 – Processus d'identification et d'apprentissage des essences de bois

cartographie des caractéristiques vraisemblables de celles-ci. Il est important de conserver visible la complexité du vivant pour maintenir la prise en compte de la variabilité et du risque d'erreur d'interprétation des observables. Par exemple, certaines espèces très proches ou similaires, anatomiquement comme dans la famille des *Dalbergia sp* peuvent être classées illégales ou légales à la commercialisation (CITES [1]).Pour ce faire, nous nous concentrons sur l'utilisation du concept de réseaux bayésiens (BNs) [12, 17] qui fournit un formalisme mathématique pratique permettant de décrire des systèmes complexes entachés d'incertitude. Ils s'appuient sur les modèles graphiques probabilistes où la structure du réseau fournit une interface intuitive à travers laquelle un individu peut modéliser des ensembles de variables en interactions et fournit une représentation qualitative des connaissances. L'incertitude relative au système est prise en compte en quantifiant la dépendance entre les variables à l'aide de probabilités conditionnelles. Les BNs permettent de combiner différentes sources de connaissances expertes avec des données expérimentales à différents niveaux et échelles de connaissances. Le choix du formalisme des BNs s'appuie sur le fait qu'ils permettent de transformer en modèle interprétable de la connaissance contenue dans des données parfois imprécises tout en intégrant une représentation du savoir incertain plus flexible que les systèmes à base de règles. Ils sont capables de propager de l'information dans toutes les directions et de porter un jugement ou d'apporter une estimation même lorsque les données ne sont pas toutes observées. Cette approche a été étudiée et utilisée dans des domaines qui vont du biomédical à la pétro-physique [23].

Ce travail présente un système d'aide à la décision basé sur des réseaux bayésiens capable d'estimer l'essence de bois présente dans une pièce de bois observée sur la base d'un questionnaire utilisant des propriétés macroscopiques visibles. Dans une optique de formation ou de situation d'apprentissage, l'outil aura la capacité de proposer la cartographie vraisemblable des propriétés macroscopiques d'une essence de bois reconnue. La figure 1 présente le squelette du pipeline capable d'effectuer ces deux types de raisonnement : proposer (1) l'espèce de bois la plus probable qui compose l'échantillon de bois observé en répondant à un questionnaire et (2) les propriétés les plus vraisemblables qui caractérisent une espèce de bois connue. Les principales contributions de ce travail sont les suivantes :

- un modèle générique compréhensible et polyvalent capable de prédire à la fois l'essence de bois dans les échantillons de bois et la cartographie de ses propriétés,
- la conception d'un outil, basé sur le réseau bayésien et intégrant dans son apprentissage des connaissances expertes issues de sources hétérogènes, utilisable en pratique (par exemple smartphone/tablette) pour :
  - aider à l'apprentissage et à la reconnaissance des essences de bois,
  - transférer les connaissances et fournir un outil attrayant pour les établissements d'enseignement et les professionnels de l'industrie du bois.
- le traitement des incertitudes dues à l'hétérogénéité du matériau bois manipulé et à la subjectivité des évalua-

1. https ://cites.org/fra

tions des utilisateurs.

Le contenu de ce document est organisé comme suit. La section 2 donne une brève introduction aux réseaux bayésiens sur lesquels la construction du modèle de ce travail est basée. La section 3 présente l'implémentation du modèle ; la section 4 met en évidence deux scénarios d'utilisation dans des contextes différents. La dernière section propose des améliorations et ouvre vers de nouvelles perscpectives.

2 Réseaux Bayésiens

Parmi les choix de modèles d'apprentissages possibles, nous avons fait celui des réseaux bayésiens en raison de leur facilité à la fois de lecture (et donc d'interprétation), mais également de leur capacité d'intégration des connaissances expertes dans leur raisonnement. Ils représentent en effet l'une de techniques les plus utilisées pour le design de systèmes de tutorat intelligents, aux côtés des systèmes à base de règles [25]. Dans notre cas néanmoins, il est important de pouvoir modéliser les incertitudes inhérentes au domaine : connaître la valeur précise d'un paramètre de décision n'est pas forcément synonyme d'une identifica-tion directe, et parfois plusieurs candidats sont à considé-rer. De plus, il s'agit d'outils se prêtant tout aussi bien à la prédiction ("A partir de ce que je sais, quel est le ré-sultat le plus probable ?") qu'au diagnostic ("Connaissant mon résultat, quelles caractéristiques seront probablement observées ?"). Deux illustrations de ces applications sont détaillées en Sect. 4.2 et 4.3.

2.1 Structure

Un réseau bayésien (RB) [12, 17] est une représentation graphique d'une distribution de probabilité multivariée qui capture des propriétés d'indépendance conditionnelle entre les variables. Formellement, un réseau bayésien est un graphe acyclique dirigé dont les noeuds représentent les variables aléatoires et dont les arcs codent les dépendances conditionnelles entre les variables. Le graphe est appelé la structure du réseau et les noeuds contenant l'information probabiliste sont appelés les paramètres du réseau. Dans un réseau Bayésien, la distribution de probabilités jointes des valeurs des noeuds peut être écrite comme le produit de la distribution de probabilité de chaque noeud et de ses parents. Si l'ensemble des noeuds parents d'un noeud X_i est désigné par $Pa(X_i)$, la probabilité jointe peut se réécrire :

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\text{Pa}(X_i)) \tag{1}$$

où $P(X_i|\text{Pa}(X_i))$ représente la probabilité conditionnelle associée à la variable X_i sachant $\text{Pa}(X_i))$. Les variables peuvent être discrètes, continues ou contenir une combinaison des deux. Dans cet article, seuls les réseaux discrets sont considérés.

2.2 Apprentissage

Différentes techniques d'apprentissage permettent d'identifier soit la structure du graphe (i.e la topologie), soit les paramètres du réseau (i.e les distributions de probabilités conditionnelles) ou une combinaison des deux à partir de données substantielles et/ou incomplètes combinées à une élicitation par des experts [9, 13, 14, 16]. Dans ce travail, la topologie du graphe sera obtenue à partir des connaissances expertes. Supposons X_i l'ensemble des variables et θ_{ijk} la probabilité que $X_i = x_k$ sachant x_j, *i.e.*

$$\theta_{ijk} = P(X_i = x_k|\text{Pa}(X_i) = x_j) \quad \begin{array}{l} i = 1, \ldots, n \\ j = 1, \ldots, c_i \\ k = 1, \ldots, r_i \end{array} \tag{2}$$

où r_i est le nombre de valeurs que la variable X_i peut prendre et c_i est le nombre de configutations de PaX_i. Baudrit et al., [2] propose un apprentissage hybride des paramètres à partir de sources de connaissances multiples (littérature, observations empiriques, avis d'experts, modèles existants, etc). L'approche consiste (1) à initialiser θ_{ij} en utilisant des distributions a priori de Dirichlet et (2) à mettre à jour successivement θ_{ij} chaque fois que des nouvelles connaissances sont disponibles et peuvent être formulée sous une forme fréquentiste. De plus, cette approche permet de pondérer l'importance des différentes sources de connaissance. C'est-à-dire :

$$\theta_{ijk}|(D_1, \ldots, D_m) = \frac{s_0\alpha_{ijk} + s_p \sum_{p=1}^{m} f_{ijk}^p}{\sum_{p=1}^{m} s_p} \tag{3}$$

où $\alpha_{ij} = (\alpha_{ij1}, \ldots, \alpha_{ijr_i})$ sont les hyperparamètres de la distribution a priori de Dirichlet pouvant être interprétés comme la taille d'une base de données virtuelle correspondant à un niveau de confiance des experts par rapport aux essais expérimentaux. D_p représente des données issues d'essais expérimentaux ou résultant de simulations ; f_{ijk}^p représente la fréquence de l'observation x_{ijk} dans la base expérimentale D_p et s_p correspond au niveau de confiance sur la source de connaissance p. En présence de données manquantes, l'algorithme EM (expectation– maximization) peut être utilisé [14, 15, 21].

2.3 Inférence

L'utilisation de ces réseaux consiste en une "requête" exprimée sous forme de probabilités conditionnelles. La tâche la plus courante consiste à estimer les probabilités marginales $P(X_Q|X_E)$ où X_Q est un ensemble de variables de requête et X_E un ensemble de variables observées. L'inférence consiste alors à calculer la probabilité de chaque état d'une variable X_Q lorsque nous connaissons l'état pris par les autres variables X_E. Pour plus de détails sur l'inférence, le lecteur peut se référer à [20] qui présentent différents types d'algorithmes d'inférence (inférence exacte et approximative) selon la complexité et la taille du réseau [4].

3 Implémentation du modèle

3.1 Identification des variables

En amont de la modélisation, un travail de définition des variables permettant la discrimination des essences de bois a été réalisé pour établir les variables et leurs interactions.

Un groupe d'expert du domaine [6] (experts professionnels, enseignants du supérieur et chercheur du domaine de l'anatomie du bois et de la reconnaissance des essences) travaille en collaboration depuis plusieurs années directement auprès d'étudiants sur l'approche et la démarche d'apprentissage du domaine. C'est donc sur la base des règles classiques et de la formalisation de leur pratique via l'usage de techniques de recueil de connaissances [3] qu'ont été établis les critères d'observations. Un second choix pratique et pragmatique a permis de limiter leur nombre en hiérarchisant leur importance. La hierarchisation est fondée sur 2 principes : leur capacité à discriminer et leur accéssibilité en pratique. Chaque noeud, à l'exception du noeud *Species*, représente les propriétés (les plus discriminantes pour les experts) capables de décrire une essence de bois. Le noeud *Origin* décrit les origines possibles des pièces de bois, à savoir soit *Tropical*, soit *Temperate*. Les feuillus sont des arbres qui ont des feuilles larges et les résineux sont des conifères tels que les pins généralement utilisés par l'industrie de la construction. Les vaisseaux (ou pores) sont les cellules qui permettent le transport de la sève vers les feuilles. Ils sont implantés longitudinalement dans le bois et peuvent être juxtaposés ou non. Le noeud *Hardwood_Softwood* prend deux valeurs à savoir soit *Hardwood* correspondant à la présence de vaisseaux (pores) visibles et abondants, soit *Softwood* dans le cas contraire. La taille des vaisseaux et leur répartition peuvent être utilisées pour classer les feuillus en trois classes. Le noeud *Morphology* représente ces trois classes : les feuillus à pores diffus, notés *Diffuse*, ont des vaisseaux de taille similaire disposés selon une distribution relativement uniforme ; les feuillus à zone initiale poreuse, notés *Zip*, présentent en début de cerne une concentration de vaisseaux beaucoup plus gros que les vaisseaux du bois final ; les feuillus à zone semi-poreuse, notés *Semi*, montrent une évolution de la taille des vaisseaux à l'intérieur du cerne. Ce noeud est ensuite lié aux noeuds *Hardwood_Softwood*, *Origin* et prend la valeur *No* lorsque l'échantillon de bois sera estimé comme étant du bois tendre. Le noeud *Transition* caractérise la croissance des cernes qui peut avoir une transition progressive ou abrupte et prend deux valeurs *Yes* pour progressive et *No* pour abrupte ; il dépend de la configuration des trois noeuds *Origin, Hardwood_Softwood, Morphology*. Le noeud *Wood_Ray* représente les rayons ligneux (rayons médullaires) qui sont caractérisés par des lignes droites continues partant du centre perpendiculairement aux cernes de croissance et prend les valeurs *Yes* pour les rayons visibles et *No* dans le cas contraire. Le noeud *Poids* représente une estimation qualitative du poids de la pièce de bois et prend quatre valeurs : *Light, Medium, Medium_Heavy, Heavy*. Le noeud *Color* prend quatre valeurs : *Blanc, Brun, Jaune, Rougeâtre*.

3.2 Elaboration de la structure et apprentissage des paramètres

Les relations de dépendances dans les RBs ne sont pas nécessairement causales. Cependant, les experts d'un domaine précis raisonnent souvent par mécanisme de causes à effets. Les liens entre les noeuds sont des dépendances conditionnelles probabilistes et l'orientation des arcs guide la manière d'interpréter la structure et précise comment circule l'information. La Fig. 2 représente la structure du modèle établie par expertise. Par exemple, l'expert sait que les bois de couleurs sombres auront tendance à avoir une densité élevée ce qui se traduit par une relation de dépendance entre les noeuds *Color* et *Weight*. L'orientation de l'arc *Color* $\mapsto$ *Weight* indique que la connaissance de la couleur apportera de l'information moins incertaine sur le poids que le poids sur la couleur. De même, la détermination de l'origine du bois et de sa dureté permet d'avoir une bonne idée de sa morphologie potentielle (comme illustré dans l'exemple d'utilisation décrit en Sect.4.2). Si l'utilisateur n'est pas obligé de remplir les informations dans l'ordre (on pourrait imaginer un cas où le paramètre *Transition* est rempli avant *Morphology*), cette structuration construite avec l'expert permet déjà d'exprimer des connaissances causales, et guide l'apprenant en lui indiquant les paramètres essentiels à déterminer pour raffiner les résultats de l'identification .

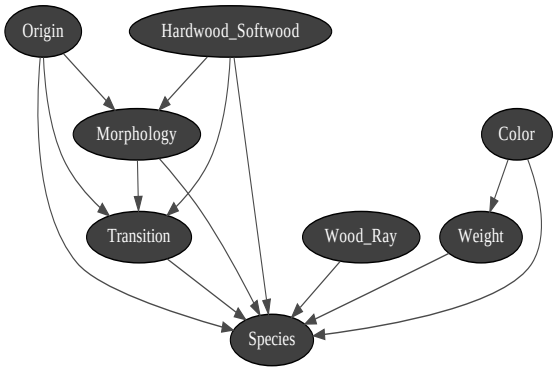


FIGURE 2 – Structure du réseau Bayésien représentant le réseau d'interaction capable de décrire les essences de bois.

TABLE 1 – Distribution de probabilité conditionnelle $P(Morphology|Origin, Hardwood_Softwood)$

		Morphology			
Hardwood_Softwood	***Origin***	***Diffuse***	***No***	***Semi***	***Zip***
Softwood	***Tropical***	0.0	1.0	0.0	0.0
	Temperate	0.0	1.0	0.0	0.0
Hardwood	***Tropical***	0.77	0.0	0.0	0.23
	Temperate	0.36	0.0	0.18	0.46

Sur la base de l'équation 3, les paramètres des réseaux bayésiens (voir Fig. 2) ont été estimés à partir des connaissances des experts et d'une base de données composée de 25 essences de bois provenant de différentes régions du monde. Le modèle a été construit et implémenté en utilisant la bibliothèque python pyAgrum [8] dédiée aux modèles graphiques probabilistes permettant aux modélisateurs de créer, gérer et effectuer des inférences avec les réseaux Bayésiens. Le tableau 1 pré-

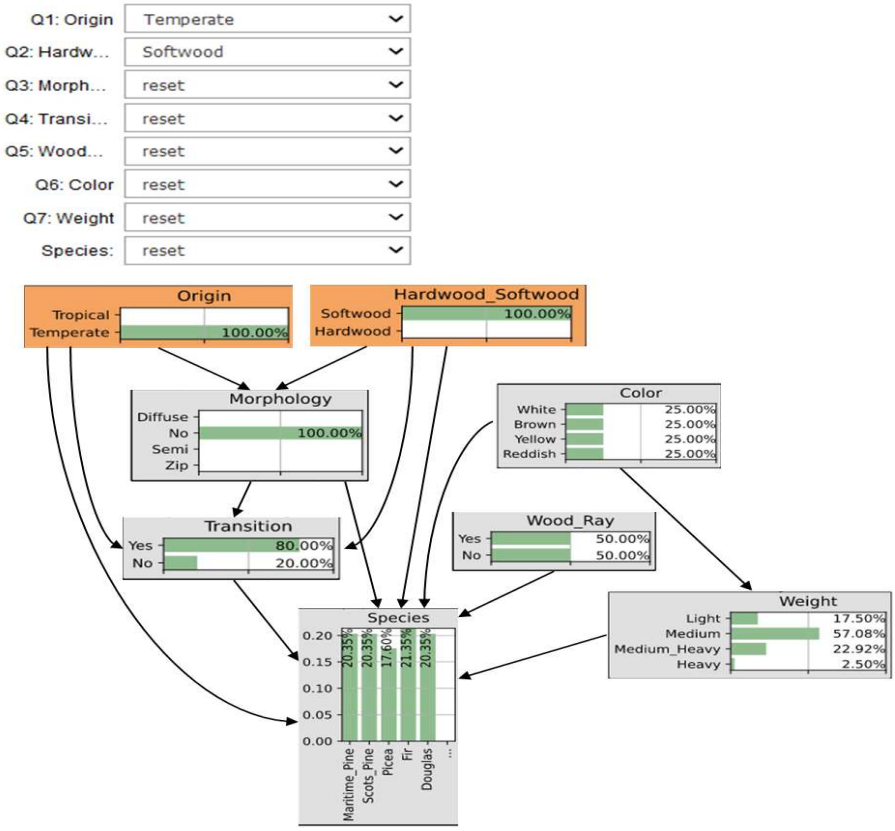


FIGURE 3 – Inférence Bayésienne sachant que la pièce de bois observée provient d'un climat tempéré et d'un bois tendre.

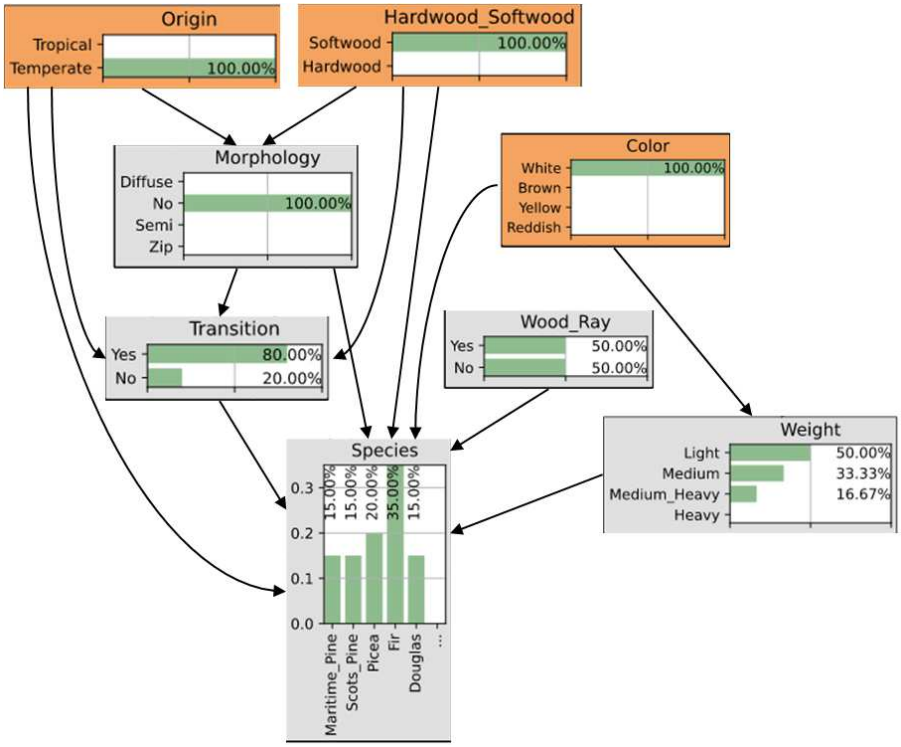


FIGURE 4 – Inférence Bayésienne sachant que la pièce de bois observée est de couleur blanche et provient d'un climat tempéré et d'un bois tendre.

sente le tableau de probabilités conditionnelles associé à la variable $Morphologie|Origine, Bois_dur_s ouple$ signifiant par exemple qu'il y a 77% de chance d'observer des vaisseaux de taille similaire disposés selon une distribution relativement uniforme dans une pièce en bois sachant que l'essence de bois utilisée provient d'une région tropicale et est issue d'un bois dur. Les probabilités $P(Morphology = No|Softwood = .)$ sont égales à 1 car les vaisseaux ou les pores ne concernent que les feuillus.

4 Utilisation et utilisabilité de l'outil

Comme indiqué en introduction, cet outil se destine à un usage éducatif, et vise à répondre à deux types de problèmes : l'identification et le diagnostic (ou appelé reconnaissance des essences dans le domaine).

4.1 Cartographie des utilisateurs

systèmeL'outil est destiné à accompagner les professionnels de la filière bois dans la pratique et se destine donc dans un premier temps à des personnes qualifiés ou en cours de qualification. Les anciens élèves ou apprenants actuels des formations du supérieur ou de la formation continue d'institut similaire à l'Ecole Supérieure du Bois sont de bons exemples. Dans un second temps, la manipulation de l'outil, son enrichissement et l'analyse des résultats est un objet de recherche pour les spécialistes de l'anatomie du bois. A terme, l'introduction d'éléments plus ergonomique et l'accès à des connaissances complémentaires, via par exemple un livre de connaissances, doit permettre un élargissement à un public peu formé voire autodidacte.

4.2 Identification

La première utilisation consiste à identifier l'essence de bois d'une pièce de bois ; les Fig. 3, Fig. 4 et Fig. 5 affichent les résultats d'inférence suite aux réponses d'un questionnaire donné sous forme électronique. La Fig. 3 montre le résultat de l'inférence étant donné que l'utilisateur répond que la pièce de bois provient d'une région tempérée et est composé de bois tendre. Le modèle élimine alors toutes les espèces de bois tropicaux et propose cinq espèces de bois possibles sans être capable de les différencier. Cela signifie que les utilisateurs doivent aller plus loin dans le questionnaire en instanciant d'autres variables comme la couleur par exemple jusqu'à ce qu'il soit possible de discriminer les espèces de bois. Une fois que le modèle a fourni sa proposition, il peut être vérifié en le comparant à une fiche d'information. La Fig. 4 montre le résultat de l'inférence étant donné que l'utilisateur a ajouté la couleur visible comme étant le blanc. Le modèle estime alors que l'essence de bois la plus probable est un sapin (un type de conifère) avec 35% de certitude (le deuxième candidat, le Picea, ayant quant à lui une certitude de 20%). De plus, le modèle estime que la transition des cernes de croissance sera probablement progressive avec une probabilité de 80%. Il reste donc une incertitude de 20% qui sera vérifiée par l'observation de l'échantillon. L'observation d'une transition abrupte à ce stade pourrait remettre en question l'origine du bois. La Fig. 5 montre le résultat de l'inférence étant donné que l'utilisateur observe finalement une transition abrupte et que le modèle met à jour la prédiction en proposant un Picea comme espèce de bois avec 40% de certitude.

4.3 Diagnostic (Reconnaissance des essences)

La deuxième utilisation de l'outil consiste à proposer une carte des descripteurs de l'essence de bois reconnue. La Fig.6 affiche le résultat de l'inférence étant donné que l'utilisateur pense observer une espèce de bois composée de bois d'Orme. Le modèle estime que la couleur du spécimen de bois devrait être brune ; l'utilisateur devrait observer des zones avec des vaisseaux beaucoup plus grands, une transition progressive des cernes de croissance, *etc*. L'étape suivante consiste à vérifier les propriétés prédites de la carte avec des échantillons de bois d'orme ou des fiches d'information sur le bois d'orme provenant d'une base de données [2] [3]. Selon les résultats de l'inférence, l'outil estime que la couleur est probablement brune avec une certitude d'environ 40% sans exclure les teintes blanches, jaunes et rougeâtres, ce qui est confirmé par les fiches Elm où la couleur est décrite comme brun clair à brun avec une teinte rougeâtre (voir Fig. 6). L'outil estime également que les utilisateurs devraient observer des zones avec des vaisseaux (pores) beaucoup plus grands que d'autres (*i.e* $P(Morphology = Zip)$ =54%), ce qui est confirmé par les caractéristiques clés de l'orme dans la fiche d'information (voir en Fig 6). Selon l'image de l'orme dans la fiche d'information, les cernes de croissance ont une transition progressive qui est bien prédite par l'outil avec une certitude d'environ 70% (voir la figure 6). La vérification entre les valeurs prédites et la fiche d'information semble confirmer que les utilisateurs sont confrontés à une pièce de bois qui provient de l'Orme. Cet exemple montre que l'outil peut être facilement utilisé sous forme de formation puisqu'il est capable de décrire les essences de bois à un niveau macroscopique. en retranscrivant sous forme de probabilités des connaissances expertes techniques. L'outil permet ainsi de simuler la potentielle variabilité des paramètres observés au sein d'une même essence, et permet une approche plus flexible qu'une base traditionnelle de règles de décisions.

5 Conclusion

Un modèle générique et un outil réaliste basé sur des graphes probabilistes ont été mis en place, permettant aux utilisateurs d'identifier à la fois (1) l'essence de bois la plus probable qui a été utilisée ou qui est présente dans un objet en bois et (2) la configuration la plus probable des propriétés de l'essence de bois reconnue. Les descripteurs choisis sont des observables, à une échelle macroscopique plus ou moins accessible selon l'état de l'échantillon de bois. Cette démarche s'inscrit dans l'approche d'apprentissage pratique et pragmatique de l'identification et de la reconnaissance des essences de bois déployée depuis 2014 à l'École Supérieur du Bois [4]. L'ajout d'essence de bois à la base de données et l'ajout de variables descriptives à

2. http ://www.woodanatomy.ch/
3. ,https ://www.fpl.fs.fed.us/research/centers/woodanatomy/
4. https ://www.esb-campus.fr/ ?lang=en

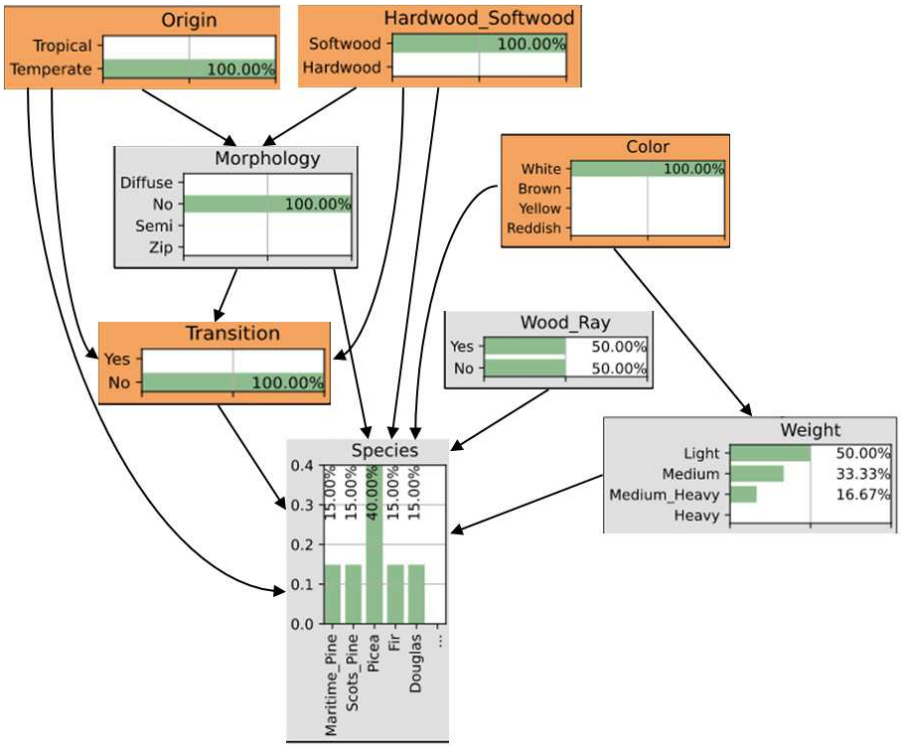


FIGURE 5 – Inférence Bayésienne sachant que la pièce de bois observée provient d'une région tempérée et de bois tendre avec une observation de couleur blanche et une transition abrupte des cernes de croissance.

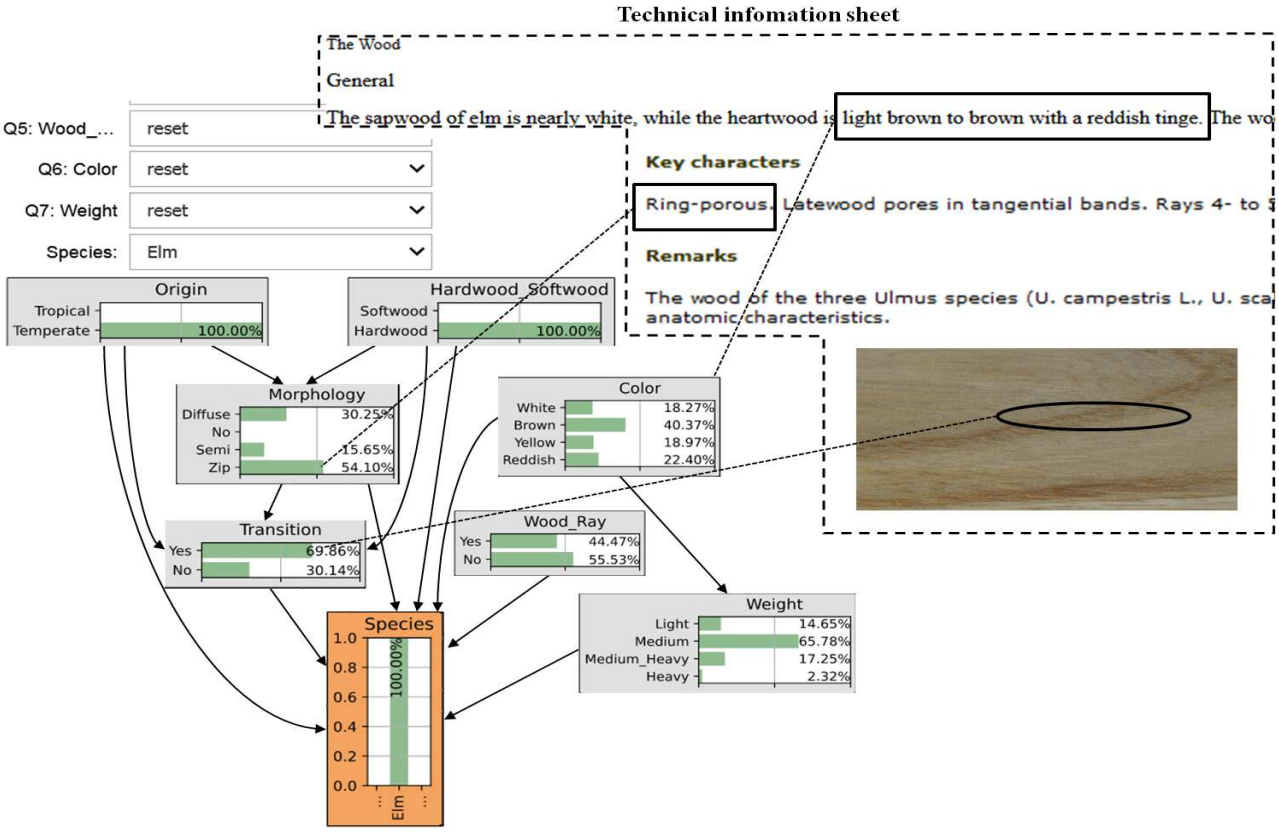


FIGURE 6 – Cartographie vraisemblable des propriétés de l'orme.

la structure du modèle permettront d'enrichir les connaissances et d'améliorer l'identification. Une amélioration de l'outil serait de relier le modèle à des bases de données de fiches techniques pour une comparaison et une vérification automatiques. Une évaluation des performances de l'outil auprès d'une base d'apprenants est également prévue, afin de valider la structure et le fonctionnement de l'application. Ces recherches sont utiles à la recherche d'informations pour les établissements d'enseignement et les professionnels de l'industrie du bois.

Remerciements

Ces travaux sont en partie financés par le programme de coopération franco-québecois Samuel-De-Champlain.

Références

[1] P., Barmpoutis, K. Dimitropoulos, I. Barboutis, N. Grammalidis, P. Lefakis. Wood species recognition through multidimensional texture analysis. Computers and electronics in agriculture, 144, 241-248, 2018.

[2] C. Baudrit, P.H. Wuillemin, N. Perrot. Parameter elicitation in probabilistic graphical models for modelling multi-scale food complex systems. Journal of food engineering, 115(1), 1-10, 2013.

[3] M. Bodineau & F. Michaud. L'apprentissage de la reconnaissance et de l'identification des essences de bois par une approche pratique et pragmatique : du recueil à la représentation des connaissances. 9èmes journées du GDR 3544 « Sciences du bois », Grenoble, France, 18-20 novembre 2020.

[4] G.F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. Artificial intelligence, 42(2-3), 393-405, 1990.

[5] P. Corbineau, N. Macchioni. LA FONCTION, D. B. D. coNNaître, recoNNaître et Nommer le bois, 2015.

[6] P. Corbineau, N. Macchioni, N. Saedlou, T. Ramananantoandro, F. Michaud. Xylothèque pédagogique, approche pragmatique de la reconnaissance et l'identification des essences de bois. 3èmes journées du GDR 3544 « Sciences du bois », Nancy, France, 12-14 novembre 2014.

[7] G. Figueroa-Mata, E. Mata-Montero, J.C Valverde-Otárola, D. Arias-Aguilar. Automated image-based identification of forest species : challenges and opportunities for 21st century xylotheques. International Work Conference on Bioinspired Intelligence (IWOBI),pp. 1-8, 2018.

[8] C. Gonzales, L. Torti, P.H. Wuillemin. Agrum : a graphical universal model framework. In Proceedings of the 30th Int. Conference on Industrial, Engineering, Other Applications of Applied Intelligent Systems, Arras, France. Springer-Verlag, 2017.

[9] D. Heckerman. A Tutorial on Learning with Bayesian Networks, In Innovations in Bayesian networks (pp. 33–82), Springer, Berlin, Heidelberg, 2008.

[10] J. Hu, W. Song, w. Zhang, Y. Zhao, A. Yilmaz. Deep learning for use in lumber classification tasks. Wood Science and Technology, 53(2), 505-517, 2019.

[11] T. He, L. Jiao, M. Yu, J. Guo, X. Jiang, Y. Yin. DNA barcoding authentication for the wood of eight endangered Dalbergia timber species using machine learning approaches. Holzforschung, 73(3), 277-285, 2019.

[12] F.V. Jensen, T.D. Nielsen. Bayesian Networks and Decision Graphs. Springer, 2010.

[13] Z. Ji, Q. Xia, G. Meng. A review of parameter learning methods in Bayesian network. In International Conference on Intelligent Computing (pp. 3-12). Springer, Cham, 2015.

[14] S.L. Lauritzen. The EM algorithm for graphical association models with missing data. Computational Statistics and Data Analysis 19, 191–201, 1995.

[15] R. M. Neal, G.E Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In Learning in graphical models. Springer, Dordrecht, 355-368 ; 1998.

[16] A. O'Hagan. Uncertain Judgements :Eliciting Experts' Probabilities,Wiley, New York,2006.

[17] J. Pearl. Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference. Morgan Kaufmann, San Diego, 552p, 1988.

[18] P. Ravindran, B.J. Thompson, R.K. Soares, A.C. Wiedenhoeft. The XyloTron : flexible, open-source, image-based macroscopic field identification of wood products. Frontiers in plant science, 1015, 2020.

[19] J.A.M Rojas, J. Alpuente, D. Postigo, I.M.Rojas, S. Vignote. Wood species identification using stress-wave analysis in the audible range. Applied Acoustics, 72(12), 934-942, 2011.

[20] A. Salmerón, R. Rumí, H. Langseth, T.D. Nielsen, & A.L. Madsen. A review of inference algorithms for hybrid Bayesian networks. Journal of Artificial Intelligence Research, 62, 799-828, 2018.

[21] B. Thiesson, C. Meek, D. Heckerman. Accelerating EM for large databases. Machine Learning, 45(3), 279-299, 2001.

[22] D.J. Verly Lopes, G.W. Burgreen, E.D. Entsminger. North American hardwoods identification using machine-learning. Forests, 11(3), 298, 2020.

[23] W. Wiegerinck, W. Burgers, B. Kappen. Bayesian networks, introduction and practical applications. In Handbook on Neural Information Processing (pp. 401-431). Springer, Berlin, Heidelberg, 2013.

[24] P. Zhao, G. Dou, G.S. Chen. Wood species identification using improved active shape model. Optik, 125(18), 5212-5217, 2014.

[25] E. Mousavinasab, N. Zarifsanaiey, S. R. Niakan Kalhori, M. Rakhshan, L. Keikha, M. Ghazi Saeedi. Intelligent tutoring systems : a systematic review of characteristics, applications, and evaluation methods. Interactive Learning Environments, 29(1), 142-163, 2021.

Une approche d'ingénierie inverse combinant ontologies et modèles relationnels probabilistes: application aux emballages bio-composites

M. Munch*[1], P. Buche[2, 3], C. Manfredotti[4], P-H. Wuillemin[5], H. Angellier-Coussy[2]

[1] I2M, U. Bordeaux, INRAE, Talence, France
[2] IATE, U. Montpellier, INRAE, CIRAD, Montpellier SupAgro, Montpellier, France
[3] LIRMM, U. Montpellier, CNRS, INRIA GraphIK, Montpellier, France
[4] UMR MIA-Paris, AgroParisTech, INRAE, Paris-Saclay University, Paris, France
[5] Sorbonne University, UPMC, Univ Paris 06, CNRS UMR 7606, LIP6, 75005 Paris, France

melanie.munch@u-bordeaux.fr

Résumé

En raison des nombreux paramètres et variables à prendre en compte, la conception de nouveau procédés de transformation pour les emballages alimentaires est un défi économique et écologique. Pour le relever, celui-ci nécessite (1) l'intégration de sources hétérogènes de données et (2) de pouvoir raisonner causalement. Dans cet article, nous présentons POND (Process and observation ONtology Discovery), un workflow dédié à l'étude de questions expertes sur des domaines modélisés par la Process and Observation Ontology (PO²), ontologie dédiée à la représentation de procédés de transformation. Nous illustrons son fonctionnement à travers un problème concret d'ingénierie inverse à partir d'une base de données inédite dans le cadre de la conception d'emballages alimentaires bio-composites.

Mots-clés

Graphe de Connaissance, Modèle probabiliste, Causalité, Emballage alimentaire

Abstract

Designing new processes for bio-based and biodegradable food packaging is an environmental and economic challenge. Due to the multiplicity of the parameters, such an issue requires an approach that proposes both (1) to integrate heterogeneous data sources and (2) to allow causal reasoning. In this article, we present POND (Process and observation ONtology Discovery), a workflow dedicated to answering expert queries on domains modeled by the Process and Observation Ontology (PO2), an ontology dedicated to represent transformation processes. The presentation is illustrated with a real-world application on bio-composites for food packaging to solve a reverse engineering problem, using a novel dataset composed of data from different projects.

Keywords

Knowledge graph, Probabilistic model, Causality, Food packaging

Acronymes

BC : Base de Connaissance ; **CD :** Contrainte Dure ; **CL :** Charge Lignocellulosique ; **CS :** Contrainte Souple ; **GE :** Graphe Essentiel ; **MRP :** Modèle Relationnel Probabiliste ; **QCc :** Question de Connaissance causale ; **RB(C) :** Réseau Bayésien Causal ; **SR :** Schéma Relationel

1 Introduction

Chaque année, l'utilisation massive de plastique résulte en une accumulation constante de déchets environnementaux, avec des conséquences désastreuses à la fois pour les écosystèmes et la santé humaine. Face à l'épuisement grandissant des énergies fossiles et de la production croissante de résidus organiques (agricoles, urbains, forestiers ou d'industries agro-alimentaires) non récupérés, des technologies innovantes ont été développées pour la production de matériaux recyclables, biosourcés et biodégradables. Parmi ces solutions, le poly(3-hydroxybutyrate-co-3-hydroxyverate) (PHBV) est un biopolymère bactérien prometteur, dégradable dans le sol et les océans, pouvant être synthétisé à partir de toutes sortes de résidus carbonés. Afin de diminuer son coût et empreinte carbone et d'agir sur le réchauffement climatique, le développement de biocomposites de PHBV avec des produits lignocellulosiques a été travaillé [7]. Néanmoins, cette augmentation dans la composition de fibres lignocellulosiques a également un impact sur la fragilité du composite final et ses aptitudes de mise-en-œuvre. Ainsi, un compromis doit être trouvé entre la quantité maximale intégrable de charge et les caractéristiques finales du bio-composite. Pourtant, trouver des liens causaux entre les différentes variables présentées depuis le jeu de données seul peut être ardu. Si les travaux précédents sur la causalité suggèrent l'usage d'interventions (i.e. changer une variable en gardant les autres constantes pour évaluer les effets) pour construire des modèles causaux [26], celles-ci peuvent de-

venir très vite chronophages et onéreuses. Dans cet article, nous présentons une alternative, POND (Process and observation ONtology Discovery), un workflow basé sur la combinaison entre une ontologie dédiée à la représentation de procédés de transformation, PO² [18], et des modèles probabilistes. L'idée principale est de combiner les connaissances expertes contenues dans le graphe de connaissance [12] avec celles prodiguées par un expert pour guider l'apprentissage d'une extension des réseaux Bayésiens (RB), le modèle relationnel probabiliste (MRP) [15]. Le modèle ainsi appris sous contraintes est alors capable de raisonner causalement sur le problème étudié, afin de répondre à des questions expertes. Les contributions originales de cet article sont (1) l'intégration complète de PO² dans un pipeline pour répondre à des questions expertes ; (2) un outil pour répondre à des questions causales permettant une approche d'ingénierie inverse ; (3) une méta-analyse de différents projets menés sur l'étude des emballages alimentaires biocomposites. La section 2 de cet article présente les prérequis utiles à la compréhension de POND, couvrant à la fois l'ontologie PO², les modèles graphiques probabilistes et la découverte causale. La section 3 introduit POND et souligne ses contributions à l'état de l'art sur la combinaison entre bases de connaissances et modèles probabilistes dans le cadre d'intégration de connaissances expertes. La section 4 illustre POND à travers l'exemple des emballages biocomposites et d'un exemple concret. Pour ce faire, nous basons notre étude sur une base de connaissances novatrice, composée à partir de différents projets.

Cet article est une version traduite d'un travail préalablement publié par les auteurs [20], et récompensé du *Best Paper Award*.

2 Travaux antérieurs

2.1 Process and Observation Ontologie PO²

PO² est une ontologie générique dédiée à la représentation des procédés de transformation. Initialement dédiée à la science des aliments [18], elle a été développée via le scénario 6 de la méthodologie NeON [30], en retravaillant une ontologie préexistante dédiée à l'éco-conception de procédés de transformation [10]. Elle a récemment été utilisée pour des produits biosourcés tels que les emballages alimentaires biocomposites. La Figure 1 présente un aperçu de ses différents concepts, décrits par 67 concepts et 79 relations. Un **procédé de transformation** est représenté ici comme une succession d'**étapes**, mises en relation par des **entités temporelles**, auxquelles sont rattachés des **résultats** expérimentaux pouvant être mesurés à de multiples échelles et unités sur différents **composants** (représentant des facteurs d'intérêts). La version 2.0 de PO² est implémentée en OWL 2 [1], et publiée sur AgroPortal [2] en licence publique Creative Commons Attribution International (CC BY 4.0) [3].

1. https ://www.w3.org/TR/owl2-overview/
2. http ://agroportal.lirmm.fr/ontologies/PO2
3. https ://creativecommons.org/licenses/by/4.0/

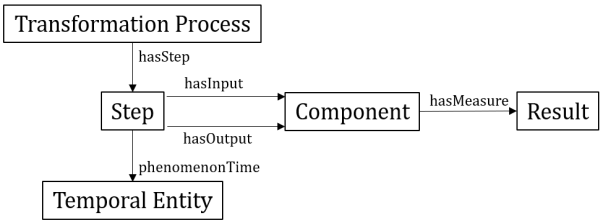


FIGURE 1 – Représentation simplifiée des composants principaux de l'ontologie PO².

2.2 Modèles Probabilistes : RB et MRP

2.2.1 Réseaux Bayésiens

Un RB est la représentation de probabilités jointes sur un ensemble de variables aléatoires dont les dépendances probabilistes sont encodées par un graphe acyclique orienté. Ainsi, chaque nœud représente une variable, et chaque flèche une relation probabiliste. L'apprentissage d'un RB se fait généralement en deux temps : la structure d'abord, puis les dépendances probabilistes. Dans notre cas, cet apprentissage se fait en utilisant l'algorithme classique Greedy Hill Climbing [5] guidé par un score BIC [27]. Un RB dont chaque relation traduit dans le domaine représenté une relation causale est appelé **RB causal** (RBC). Un exemple de RBC est présenté par la Figure 4, tandis que la Table 2 montre un exemple de dépendance probabiliste sous la forme d'un tableau de dépendance conditionnelle (dans ce cas, l'évolution des probabilités de la variable **Contrainte à la rupture** en fonction des valeurs de la variable **Charge**).

2.2.2 Graphes Essentiels

Pour chaque RB, il est possible de déduire le graphe essentiel (GE), graphe acyclique semi-orienté exprimant la classe d'équivalence de Markov du réseau. Bien que RB et GE partagent la même structure (mêmes nœuds et mêmes liens), certaines relations ne sont pas orientées dans le GE, se traduisant par une arête sans flèche. Cette présence (ou non) d'orientation traduit la nécessité de la conserver pour préserver les relations d'indépendances encodées dans le graphe. Plus généralement, si une relation n'est pas orientée dans le GE, alors celle-ci peut être inversée dans le RB sans modification des relations de dépendance sous-jacentes ; en revanche, si elle est orientée, alors son inversion nécessite le réapprentissage de toute la structure du RB. La Figure 2 illustre deux exemples de classes d'équivalence de RB et leur GE associé. Comme nous le verrons par la suite, le GE est donc une source d'information pouvant être utilisée, dans un certain cadre (que nous définirons par la suite) pour la découverte causale.

2.2.3 Modèles Relationels Probabilistes

Le MRP est une extension orientée objet des RB. Là où un RB ne nécessite qu'une seule couche d'information pour être décrit, le MRP en nécessite deux : (1) le schéma relationnel (SR), qui fournit une description qualitative des classes et de leurs variables (appelées dans ce cadre attributs) composant sa structure ; et (2) le modèle relationnel

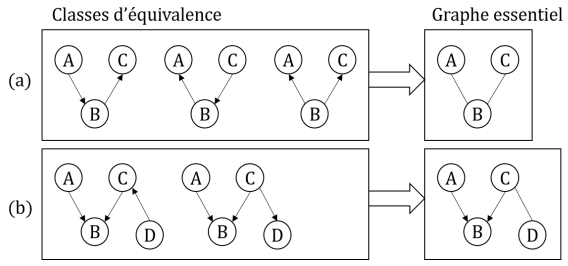


FIGURE 2 – Exemples de deux classes d'équivalence de Markov et leur graphe essentiel attribué. (a) Dans le premier exemple, les arcs des RB peuvent être orientés dans toutes les directions sans incidence sur les relations entre variables : le GE est donc complètement non-orienté. (b) Dans le second en revanche les trois variables A, B et C forment une structure particulière ne pouvant être modifiée : elle est donc marquée dans le GE.

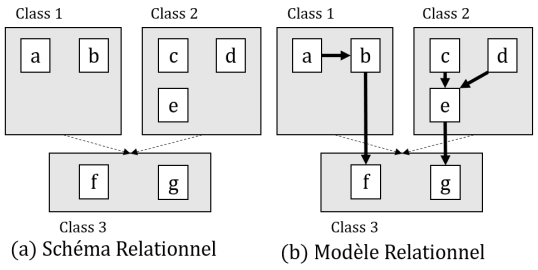


FIGURE 3 – Exemple d'un MRP et de la définition de ses (a) hautes et (b) basses structures.

(MR), qui contient toutes les informations quantitatives sur la distribution probabiliste entre les attributs. Deux classes peuvent être liées dans le SR par des liens relationnels, qui indiquent la direction des liens probabilistes entre les attributs : ainsi, dans la Figure 3, le lien de **Classe 1** vers **Classe 3** dans le SR force l'orientation de la relation entre les attributs b et f dans le MR. En revanche, en l'absence de lien relationnel (comme par exemple entre **Classe 1** et **Classe 2**), aucun lien ne peut être appris. Enfin, l'orientation des liens au sein d'une même classe n'est pas forcéee. L'apprentissage entièrement automatique d'un MRP, en deux temps (pour le SR puis le MR) est complexe : dans notre cas, nous construisons à la main le SR, réduisant la complexité d'apprentissage du MR à celle d'un RB [15]. Une fois le SR et le MR définis, le MRP est similaire à un RB.

2.2.4 Apprentissage sous contraintes

L'apprentissage sous contraintes (*i.e.*, guidé par des connaissances restreignant l'espace de recherche) dans le cadre des RB permet d'améliorer grandement la précision des résultats, qu'il s'agisse de celle de la structure [9] ou des paramètres [8]. Cela se vérifie d'autant mieux dans le cadre de petites bases de données [23]. Dans cette optique, des travaux précédents ont proposé des méthodes pour imposer un ordre total [6] ou partiel [25] sur les nœuds, créant ainsi des contraintes directionnelles sur leurs relations. Dans notre cas, la définition manuelle du SR permet de créer un ordre partiel entre les variables. Celui-ci transcrit les connaissances expertes apportées par la base de connaissance (BC) et l'expert, favorable à un contexte de découverte causale comme présenté dans [21]. Puisque dans notre cas nous intégrons des connaissances expertes lors de l'apprentissage, celui-ci est considéré comme étant fait sous *contraintes causales*.

2.3 Découverte de Connaissances

2.3.1 Découverte Causale

Il est bien connu qu'une corrélation n'implique pas forcément de causalité : il est donc important de répondre à certains critères précis lorsque l'on se place dans un contexte de découverte causale. Parmi les principaux, on retrouve l'absence de facteurs extérieurs non mesurés (la **suffisance causale** [29]) ; et l'absence de données erronées, de biais de sélection ou de cas déterministique [16]. D'une façon générale, le jeu d'apprentissage doit être représentatif du domaine que l'on veut modéliser : si un évènement n'est pas représenté, ou si les proportions ne sont pas conformes à la réalité (par la sur-représentation d'un cas par exemple), il devient alors impossible de tirer de bonnes conclusions causales. La découverte de causalité à partir d'un jeu de données peut se faire par des calculs d'indépendance entre variables [29, 31] ; ces méthodes néanmoins ne permettent pas d'intégrer de connaissances extérieures. Certains travaux ont proposé l'utilisation du GE pour apprendre des modèles causaux : [17] propose deux stratégies optimales pour suggérer des interventions afin d'apprendre des modèles causaux ; [28] et [4] utilisent le GE pour construire un RBC tout en maintenant un nombre limité d'interventions possibles. Tout comme les méthodes à base de calculs d'indépendances, ces méthodes n'intègrent pas de sources de connaissance extérieures. Notre but étant au contraire de pouvoir intégrer les connaissances expertes contenues dans des sources externes, nous avons donc choisis de combiner l'apprentissage avec des ontologies.

2.3.2 Combinaison avec les Ontologies

La découverte causale à partir de données seules est une tâche ardue. Pour cette raison, de nombreux travaux ont entrepris d'intégrer les connaissances contenues dans les ontologies pour apprendre des modèles probabilistes et découvrir de nouvelles relations. Par exemple, différentes extension d'ontologies permettent l'intégration directe de raisonnement probabilistes (comme BayesOWL [11, 32], ou HyProb-Ontology [19]). Si elles permettent de raisonner, elles n'autorisent néanmoins pas l'apprentissage de nouvelles relations. Au contraire, d'autres travaux partent de la structure de l'ontologie pour construire un RB, en considérant par exemple les propriétés objets comme des relations probabilistes [14] ou causales [1] ; cet a priori ne peut néanmoins pas s'appliquer à PO², où de nombreuses relations ne peuvent être directement traduites causalement. D'autres méthodes, finalement, ont été développées pour répondre à certains cas précis, ne pouvant être étendus à d'autres cas : ainsi, [2] utilise des modèles prédéfinis pour

réaliser des diagnostics médicaux, qui ne peuvent être étendus à d'autres applications médicales. Il est à noter que bien que POND se base sur une unique ontologie (PO²), la complexité de celle-ci lui permet d'aborder de nombreux problèmes sur des applications plus larges qu'une simple ontologie de domaine. Dans notre cas, il est ainsi possible de raisonner sur n'importe quel procédé de transformation, pour peu qu'il soit descriptible par l'ontologie.

3 POND : PO² Ontology Discovery

Dans cette section nous présentons POND, dont le but est l'intégration de connaissances expertes dans l'apprentissage d'un modèle probabiliste afin de raisonner dessus. Nous nous concentrerons ici sur les différentes sources pouvant être utilisées pour répondre à de complexes questions probabilistes et causales. Notre application finale étant un problème d'ingénierie inverse, nous mettrons un accent particulier sur la découverte causale et les applications possibles offertes par cette dernière.

3.1 Intégration de Connaissances

Expression. Les connaissances expertes peuvent venir : (1) de données expérimentales, récupérées auprès de différentes sources (telles que des publications, livres ou données produites au sein de projets) ; (2) d'entretiens directs sollicités auprès d'experts du domaine. La plupart de ces informations peuvent ensuite être directement structurées dans l'ontologie PO² : cela concerne les données factuelles, descriptives du domaine. L'intérêt de cette sémantisation est qu'elle permet d'établir un vocabulaire cohérent et complet, indispensable pour la suite. À partir de celui-ci, l'expert peut ainsi formuler des **Questions Expertes** de deux types : certaines restent à un niveau descriptif, et peuvent être répondues en requêtant directement l'ontologie (**Questions de Compétences**) ; d'autres requièrent en revanche un raisonnement probabiliste, et nécessitent l'apprentissage d'un modèle (**Questions de Connaissance**). Dans cet article, nous nous focaliserons sur les questions de connaissance causales (QCcs), qui peuvent être formalisées de deux façons, en définissant X_i et X_j deux groupes de variables du domaine :

QCc_1 Est-ce que X_i a une influence sur X_j ?
QCc_2 Quel est l'impact de X_i sur X_j ?

Ces deux questions illustrent la double lecture offerte par les RBC : alors que QCc_1 se concentre sur l'aspect descriptif des relations apprises (pouvant être déduit directement du graphe), QCc_2 interroge plutôt leur nature (pouvant être analysée à partir des tables de probabilités conditionnelles).

Intégration. Une fois la QCc définie, le modèle peut être construit. Comme décrit en Section 2.2, le but ici est de transcrire les connaissances expertes exprimées au préalable dans le SR du MRP. L'originalité de notre approche repose sur la façon dont cette connaissance est intégrée :

1. **Par l'alignement des variables de l'ontologie dans le SR.** Grâce au vocabulaire commun défini dans PO², l'expert peut facilement extraire les variables intéressantes pour sa question. La définition sémantique permet également de récupérer de façon automatique les valeurs associées, même si elles ont été mesurées dans différents contextes, et les unifier (ou non). Cette question de l'unification est posée par la structure même de certains procédés. Ainsi, supposons un procédé de transformation où, dépendant des itérations, une température donnée est mesurée soit à l'étape A, soit à l'étape B. Dans ce cas, l'expert peut alors décider si ces températures sont similaires (i.e. elles peuvent être comparées, et donc unifiées), ou au contraire si elles représentent deux mesures différentes. Ce genre de problème ne peut être résolu par l'ontologie directement : dépendant de la QCc, la distinction entre les températures des étapes A ou B peut être pertinente ou non. La structuration sémantique du vocabulaire est donc ici importante : elle permet à l'expert de construire le modèle, tout en utilisant un vocabulaire qui lui est accessible, pour décrire des connaissances que seul lui peut fournir. À partir de PO², l'expert peut alors spécifier les attributs à représenter dans le SR, en spécifiant à chaque fois l'itinéraire, l'étape et le composant sur lequel a eu lieu la mesure. Cela permet de lier à chaque variable ses valeurs, à savoir ici les datatypes, qui permettent de composer la base d'apprentissage du RBC.
2. **Par la définition des contraintes de précédence.** Par définition, les relations entre classes du SR établissent les contraintes de précédence : si une telle contrainte existe entre la classe comprenant la variable A vers la classe comprenant la variable B, alors un lien appris sera forcément orienté de A vers B. Au contraire, deux variables définies dans une même classe n'auront aucune contrainte de direction quant à leur potentielle relation. Ces contraintes de précédences traduites en liens relationnels entre classes peuvent être déduites des informations temporelles contenues dans le BC (par exemple, une mesure faite à une étape au temps t peut avoir une influence sur une mesure faite à une étape au temps $t + n$, mais pas l'inverse). Elles peuvent être également fournies par l'expert, en tant que causalités potentielles du type "Je sais que la variable A peut expliquer la variable B" ou, au contraire, "A et B ne peuvent pas partager de lien."

Notre contribution dans cette section est la formalisation de l'intégration de connaissances dans un workflow : grâce à PO², tout procédé de transformation peut être aisément intégré dans un SR, grâce au vocabulaire commun défini avec l'expert. Ce dernier permet également la constitution automatique de la base d'apprentissage (utilisée lors de la définition du MR) directement à partir des faits contenus dans la BC.

3.2 Validation Causale

Une fois le SR construit par l'expert par l'intégration de ses contraintes de précédences, le MR peut être appris ; il devient alors possible d'instancier le MRP (défini par le SR et

le MR) en un RB. Dans notre cas, nous considérons que les connaissances expertes intégrées ont permis d'apprendre le modèle sous contraintes causales, permettant ainsi d'utiliser le RB pour déduire des connaissances causales [22]. Cela est dû au fait que nous considérons ce modèle appris comme l'intersection entre (1) toutes les contraintes contenues dans le jeu de données utilisé par l'apprentissage (exprimé par le GE) ; et (2) toutes les connaissances expertes intégrées dans l'apprentissage (exprimées par le SR). Encore une fois, il est important de rappeler que ces déductions se basent sur le fait que nous nous considérons en contexte favorable à la découverte causale tel que décrit en Section 2.3.1. La validation se déroule alors de la façon suivante :

— Si une relation est apprise entre deux variables ayant une contrainte de précédence définit par l'expert, alors la causalité est validée par la connaissance experte.
— Si une relation apprise est représentée par un arc orienté dans le GE, alors la causalité est validée par les données à travers le GE. Cette déduction suit le raisonnement suivant : si la base de faits utilisée pour l'apprentissage est fiable, alors cette relation ne peut être orientée que de cette façon pour respecter les contraintes appliquées lors de l'apprentissage et les indépendances calculées entre les variables inhérentes à la base d'apprentissage. De plus, cette relation est validée même si aucune contrainte de précédence n'a été placée par l'expert entre ces variables.
— Si une relation est apprise mais qu'elle n'est validée ni par l'expert ni par le GE, alors il est impossible d'en déduire de la causalité.

Idéalement, cette découverte causale a pour but de valider causalement toutes les relations du RB, permettant ainsi de définir un RBC. Néanmoins, même si toutes les relations ne peuvent être validées, elle permet également :

— **D'aider l'expert à critiquer.** Puisque nous cherchons à modéliser des domaines réels, l'évaluation directe est parfois très compliquée (voire impossible) à réaliser directement. En revanche, en présentant les relations causales apprises à l'expert, nous lui donnons un outil pour les critiquer et les questionner à partir de ses connaissances propres, en suggérant par exemple de nouvelles hypothèses à vérifier expérimentalement.
— **De répondre aux QCcs.** Une QCc dépend de la découverte causale pour être résolue : QCc_1 regarde la présence (ou l'absence) de relation entre les variables concernées, tandis que QCc_2 utilise ces relations pour raisonner sur les tables de probabilités conditionnelles. Si les relations nécessitant d'être vérifiées ne sont pas validées, alors il est impossible de répondre aux QCcs dans un cas comme dans l'autre.

Il est important de noter qu'en cas de non-validation, plusieurs solutions sont possibles : l'expert peut fournir de nouvelles connaissances (sous la forme de données supplémentaires, ou de nouvelles contraintes de précédences) ; de nouvelles expériences peuvent être suggérées, pour par exemple compléter des trous de connaissance identifiés ; ou bien la QC peut être redéfinie. Si néanmoins aucune de ces solutions ne fonctionne, alors la BC est jugée insuffisante pour répondre aux questions actuelles de l'expert.

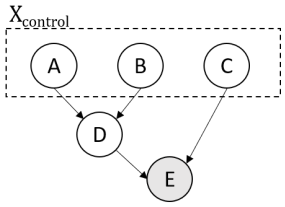


FIGURE 4 – Exemple d'un court RBC. L'ensemble $X_{control}$ représente les variables de contrôle, i.e. les variables sur lesquelles l'expert peut intervenir. E représente la variable cible.

Une application de ce protocole de validation est donné en exemple en Section 4.2.2.

3.3 Inférences Causales

Nous avons vu jusqu'à présent comment intégrer les connaissances expertes pour apprendre un modèle, et comment valider celui-ci. Si cela est généralement suffisant pour répondre directement aux questions de type QCc_1 (en vérifiant la présence ou absence de relations validées causalement), les QCc_2 et leurs dérivées nécessitent une analyse plus approfondie. Pour illustrer ce besoin, nous considérerons le RBC présenté en Figure 4, et supposerons qu'il a été validé causalement. Nous considérons également la QCc QCc_{ex} : *"Quelle intervention dois-je faire pour maximiser la variable E ?"*, combinaison entre la QCc_1 (*"Quelles variables ont un impact sur E ?"*) et la QCc_2 (*"Quelle est l'influence de ces variables ?"*).

Afin de répondre à QCc_{ex}, nous devons tout d'abord identifier les variables sur lesquelles il sera possible d'intervenir. Dénotées comme *variables de contrôles*, elles se distinguent des variables ayant également un impact sur la variable cible, mais sur lesquelles aucune intervention n'est envisageable. Dans notre exemple, nous pouvons ainsi voir que les parents directs de E sont D et C. Néanmoins, D n'appartient pas à $X_{control}$: il est donc nécessaire de remonter jusqu'à ses parents, A et B. Nous définissons donc pour la variable E l'ensemble $X_{inter} = X_{control} \cap Parents(E)$ (dans notre cas, $\{A, B, C\}$), qui contient les variables sur lesquelles on peut intervenir pour répondre à QCc_{ex}. En effet, puisque nous considérons un RBC, alors intervenir sur un parent aura un effet sur les enfants. En pratique, cela signifie que pour chaque combinaison de valeurs de A, B ou C, les valeurs de E et leurs probabilités seront affectées : cela constitue une base de tous les scénarios potentiels, dont il faudra extraire celui permettant de répondre à QCc_{ex}. Pour faire cela, l'expert est à nouveau sollicité pour définir ses propres critères d'acceptabilité, tels que "Quelles valeurs sont préférables pour la variable cible ?", ou "Quelles conditions devraient s'appliquer sur X_{inter} ?".

Ces critères sont classifiés en deux sortes :

— **Critères durs.** Certaines valeurs ou combinaisons de valeurs sont impossibles à atteindre : les scénarii correspondants sont alors automatiquement retirés. Par exemple, les experts peuvent souhaiter que la somme des variables de X_{inter} ne dépassent pas un certain seuil ; ou bien ils souhaitent exclure certaines valeurs de E. Dans notre cas, puisque l'on souhaite maximiser E, alors on ne considère pas ses valeurs les plus basses.

— **Critères souples.** Dans certains cas, l'expert a besoin de classer ses préférences dépendant du contexte. Peut-être qu'avoir une haute valeur de E n'est pas intéressant si A est également élevée ; ou peut-être qu'une valeur plus basse de E, mais avec une plus haute probabilité de réalisation, est un cas plus intéressant qu'un meilleur scénario n'ayant aucune chance d'arriver.

Définir ces critères permet de mieux définir les besoins de l'expert, et ainsi trouver la réponse à QCc_{ex} y correspondant le mieux. Nous montrerons en Section 4.3 comment cette approche peut être appliquée concrètement dans le cadre d'un problème d'ingénierie inverse.

4 Application aux Emballages Biocomposites

Pour la suite, nous définissons la QCc_{bio} : "Quelles caractéristiques de la charge ligno-cellulosique permettent d'optimiser les propriétés mécaniques de l'emballage ?".

4.1 Présentation de la Base de Connaissance

Nous avons collecté des données de cinq projets tournés vers le développement de biocomposites à partir de PHBV et de charges lignocellulosiques (CLs) provenant de déchets organiques tels que des rejets de cultures (*Chercheur d'avenir région Languedoc-Roussilon MALICE* et *H2020 NoAW*), des dérivées de l'industrie agro-alimentaire (*FP7 EcoBioCAP*) ou de déchets urbains (*H2020 Resurbis*). Les CLs de ces projets ont été obtenues par fractionnement à sec de la biomasse pure. Enfin, pour comparer, des fibres de cellulose pure ont également été utilisées comme références dans le cadre du projet *H2020 Usable*. Au final, cela constitue une base d'apprentissage de 88 formulations décrites par 15 attributs différents [24].

4.2 Intégration des Connaissances Expertes

L'intégration des connaissances expertes se déroule en deux temps : (1) la correspondance des attributs intéressants de la BC jusqu'au SR, et (2) la définition des contraintes de précédence potentielles. Dans cette section, nous présentons les résultats principaux utilisés pour apprendre le modèle final, ainsi qu'un exemple d'intégration de critiques expertes.

4.2.1 Définition du SR

Sélection des Attributs.[4] L'expert décrit une CL par trois catégories d'attributs : la composition biochimique (**cellulose**, **hémicellulose**, **cendres** et **lignine**) ; le diamètre médian apparent (**D50**) ; la **charge de remplissage** (i.e. la quantité ajoutée au produit final). Le produit final, quant à lui, est décrit par quatre catégories distinctes d'attributs : les propriétés mécaniques (**contrainte à la rupture**, **stress à la rupture** et **module de Young**), la **perméabilité** (à la vapeur d'eau), les propriétés thermiques (températures de **cristallisation** et de **fusion**), et la dégradation thermale (températures de **début** et de **pic**). Parmi ces dernières catégories, seules les propriétés mécaniques sont nécessaires pour décrire QCc_{bio} ; néanmoins, dans un contexte de découverte causale (pas d'attributs manquants) et dans une optique de facilitation des critiques expertes, nous intégrons dans un premier temps les autres catégories.

4. Pour le reste de l'article, tous les attributs représentés dans le modèle seront indiqués en **caractères gras**.

Lignine	[0 ;19] *(32)*	]19.4 ;26.4] *(30)*	]26.4 ;49] *(23)*	
Charge	[2 ;4] *(10)*	]4 ;11] *(34)*	]11 ;21] *(22)*	]21 ;50] *(19)*
CR	[0.2 ;0.5] *(19)*	]0.5 ;0.8] *(44)*	]0.8 ;1] *(15)*	]1 ;1.07] *(3)*

TABLE 1 – Extrait de la discrétisation utilisée dans notre application pour la **Lignine**, la **charge de remplissage** et la **contrainte à la rupture** (CR) *(quantité de données pour une catégorie).*

Discrétisation des Attributs. Puisqu'il s'agit pour la plupart de mesures réalisées sur les produits, les données contenues dans notre BC sont pour la plupart continues (i.e. elles ne peuvent être automatiquement rangées dans des catégories distinctes discrètes). Cependant, de par leur nature, les RB classiques ne peuvent apprendre à partir de ce type de données ; il est donc nécessaire de passer par une phase de discrétisation des variables considérées. Cette étape est importante, car pouvant influencer l'apprentissage des différentes relations entre les variables et ainsi changer l'interprétation du modèle. Elles sont ainsi très sensibles aux retours prodigués par les experts : il est donc important de convenir d'une discrétisation proposant une description équilibrée des classes (pas de sur-représentation d'une valeur pouvant déséquilibrer le modèle), mais présentant également une cohérence avec le domaine cible. Dans notre cas, nous cherchons par exemple à déterminer si les caractéristiques cibles sont dégradées (ratio valeur finale sur valeur initiale strictement inférieure à 1), conservées (valeur égale à 1) ou améliorées (valeur strictement supérieure à 1). Un exemple de la discrétisation appliquée dans notre application est donnée dans la Table 1 : les variables de contrôle sont réparties de façon équilibrées par rapport au jeu de données initial (exemple de la **lignine**), tandis que les variables cibles ont une discrétisation choisie par l'expert (exemple de la **charge** ou de la **contrainte à la rupture**).

Définition des Contraintes de Précédence. Dans notre cas, l'expert définit initialement deux contraintes de précédence qui seront précisées par la suite :

— Entre les variables du CL et les caractéristiques finales de l'emballage. Ainsi, les premières sont considérées comme des variables de contrôle pouvant avoir un impact sur les secondes, qui décrivent le résultat final. On définit donc deux classes dans le SR, avec un lien relationnel allant de la classe des

variables de contrôle vers la classe des variables de description des caractéristiques.

— Entre les différentes catégories des variables de description des caractéristiques. Cette distinction permet de considérer chaque catégorie comme un sous-groupe indépendant des autres (e.g., les caractéristiques mécaniques n'ont aucune influence sur les caractéristiques thermales). Pour modéliser ceci, la classe définie précédemment est elle-même compartimentée en différentes sous-classes indépendantes les unes des autres.

4.2.2 Retour Expert

Une fois le premier modèle appris, une discussion avec l'expert est requise pour critiquer à la fois (1) les relations apprises et (2) les dépendances probabilistes. En illustration, nous considérons le modèle appris présenté en Figure 5. Dans cette situation, l'expert a relevé plusieurs incohérences :

— La somme des constituants de la CL vaut 100 : il est donc cohérent que des corrélations soient apprises entre eux. Néanmoins, celles-ci n'ont aucun sens d'un point de vue causal. Par conséquent, il est décidé de les placer dans des sous-classes indépendantes : cela traduit le fait que l'expert a un total contrôle sur elles et peut aidément les faire varier de façon indépendante. Ce choix de modélisation conduira en Section 4.3 à l'élaboration de la contrainte CD_1 garantissant qu'une formulation soit techniquement possible (i.e. la somme des constituant vaut toujours 100).

— La température de **fusion** ne peut pas expliquer la température de **cristallisation** : la relation apprise est une corrélation, pas une relation causale. Pour y remédier, les deux paramètres sont séparés dans des sous-classes indépendantes.

— La **contrainte à la rupture** n'est contre toute attente expliquée par aucun paramètre. Une nouvelle discrétisation est testée pour tenter de mieux représenter la variable.

Le retour expert permet d'identifier des trous de connaissances (i.e. des cas non représentés dans la BC pouvant entrainer des apprentissages incomplets). Ainsi, le modèle décrit montre que lorsque la **charge de remplissage** $\in$]21 ;50[, alors la température de **fusion** $\notin$]1 ;1.02[. Cela peut être dû à deux raisons : (1) il s'agit bien d'un trou de connaissance, et la BC doit être complétée ; (2) il s'agit d'un cas normal, ne nécessitant pas d'ajouts de données supplémentaires. Dans cet exemple, l'expert a bien pu confirmer que l'absence d'amélioration de la température de **fusion** était cohérent dans le cas d'une augmentation de la **charge**.

Après quelques allers-retours, un RS a été retenu pour l'étude de la QCc. Celui-ci est présenté en Figure 6.

4.3 Résolution de la Question de Connaissance

Nous considérons maintenant un RBC entièrement validé et permettant de répondre à la QCC, présenté en Figure

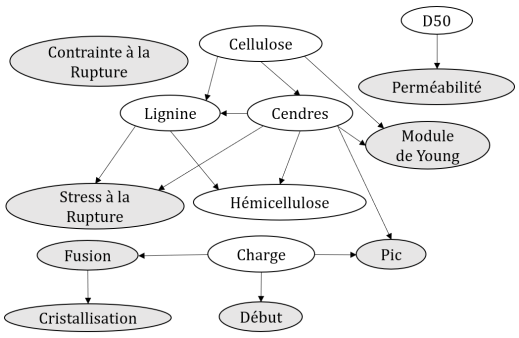


FIGURE 5 – Modèle appris après une itération et critiqué par l'expert en Section 4.2. Les variables de contrôles sont indiquées en blanc, et les variables à expliquer en gris.

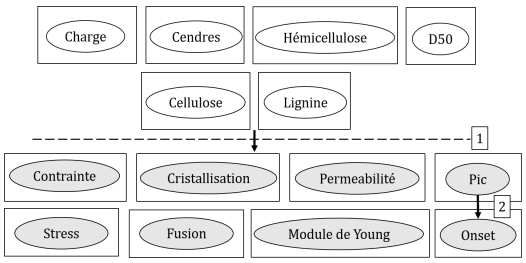


FIGURE 6 – Schéma relationnel finalement retenu après validation experte. Chaque rectangle représente une classe du SR, tandis que chaque ovale représente une variable. Deux types de liens relationnels sont indiqués : (1) indique le lien depuis toutes les classes au dessus de la ligne vers celles en dessous ; (2) indique le lien établi depuis la température de **pic** vers celle du **début**.

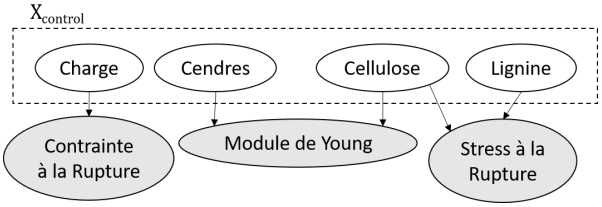


FIGURE 7 – Extrait du RB final sélectionné pour répondre à QCc_{bio}. Puisque toutes les relations représentées sont influencées par des contraintes de précédence expertes, on considère ce modèle *causal*.

	Contrainte à la rupture			
Charge	]0.24 ;0.5]	]0.5 ;0.8]	]0.8 ;1]	]1 ;1.07]
]2 ;4]	0.0076	0.4924	***0.4924****	0.0076
]4 ;11]	0.002	***0.770****	0.1620	***0.0660****
]11 ;21]	0.3624	0.4522	0.1826	0.0028
]21 :50]	***0.5747****	0.2630	0.1071	0.0552

TABLE 2 – Probabilités conditionnelles pour la **contrainte à la rupture**. (***Vraisemblance maximale****).

7, appris à partir du RS présenté en Figure 6. Dans une optique de simplification, nous présentons ici une version simplifiée où toutes les variables non pertinentes ont été retirées. QCc_{bio} nécessite deux interventions pour maximiser l'amélioration des propriétés mécaniques considérées : (1) la **charge de remplissage** et (2) la composition du CL.

4.3.1 Charge Optimale

D'après la Figure 7, la **charge** a un impact sur les propriétés mécaniques uniquement à travers la contrainte à la rupture. À partir de la table des probabilités conditionnelles (Table 2), plusieurs lectures sont possibles dépendant des critères experts élicités :

— Si l'on recherche la plus haute valeur possible de la **contrainte à la rupture** (]1 ;1.07]), les probabilités sont toutes quasi-nulles. Il s'agit donc d'un critère non envisageable réalistiquement.

— En fixant un critère dur visant la deuxième meilleure valeur de la **contrainte à la rupture** (entre 0.8 et 1), une **charge** $\in$]2 ;4] pourrait être considérée : elle garantit en effet une probabilité de 0.49, signifiant qu'en moyenne un produit final sur deux atteindra la valeur souhaitée (et obtiendra une valeur comprise entre 0.5 et 0.8 dans la plupart des autres cas).

— Dans le cas d'un procédé industriel, en revanche, l'expert pourrait souhaiter placer un critère dur sur la probabilité de réussite plutôt que sur la valeur cible, afin de garantir une stabilité des résultats. Dans ce cas, il semble plus raisonnable de considérer une **charge** $\in$]4 ;11], garantissant une probabilité de 0.77 d'obtenir une **contrainte** $\in$]0.5 ;0.8].

4.3.2 Prospections de CLs

D'après le RB présenté en Figure 7, le **module de Young** et le **stress à la rupture** dépendent de la composition du CL considéré (de la teneur en **cendres**, **cellulose** et **lignine**). Dans cette partie, nous utilisons ce résultat pour proposer de nouvelles CLs permettant de maximiser les valeurs de ces paramètres. Pour cela, nous cherchons dans @Web [3], une BC contenant les informations de composition de nombreuses biomasses [13], afin d'en proposer de nouvelles pertinentes. D'une façon similaire au paragraphe précédent, nous introduisons de nouveaux critères d'acceptabilité durs (CD) et souple (CS) :

CD_1 Pour proposer de nouvelles CLs réalistes, il est important que la somme de ses constituants ne dépasse pas 100 (en d'autres termes, la biomasse simulée doit être possible). En fixant $x \in$ {**Cendres**, **Cellulose**, **Lignine**, **Hémicellulose**} et l'intervalle associé $[x_{min}; x_{max}]$ déterminé par la discrétisation, nous fixons CD_1 tel que $\sum_x x_{min} < 100$.

CD_2 Nous voulons que les valeurs cibles soient intéressantes : on fixe donc **Module de Young** > 0.8 $\cap$ **Stress à la rupture** > 0.8.

CD_3 La probabilité de réalisation doit être supérieure à 0.25.

CS_1 Dans le cas où aucune CL potentielle n'est trouvée, nous voulons étendre la recherche à de potentiels candidats proches de la composition cible. Afin d'évaluer la qualité de tels substituts, nous définissons un score de qualité. Soit une CL m contenue dans @Web, sa composition x_m et un intervalle cible $[x_{min}; x_{max}]$ déterminé par le RB ($x \in$ {**Cendres**, **Cellulose**, **Lignine**}). On définit le score $S_m = \sum_x \sigma(m, x)$ avec $\sigma(m, x) = min(abs(x_m - x_{min}), abs(x_m - x_{max}))$.
Plus S_m est bas, plus la CL proposée est proche de la recommandation.

La requête effectuée sur @Web retourne quinze résultats, présentés partiellement dans la Table 3. Chacun de ces scénarios évalue la probabilité de succès de CD_2, sachant que l'on respecte CD_1 et CD_3. Parmi les deux scénarios présentés dans cet article, le plus probable ($p = 0.99$) n'est pas une correspondance exacte : la biomasse s'approchant le plus est l'écorce de pin, avec un S-score de 5.26 (dû notamment à son taux de cendres trop bas par rapport à la recommandation). Le second scénario présenté, plus bas en terme de probabilité de réalisation ($p = 0.82$), est quant à lui une correspondance parfaite avec l'enveloppe de riz. Malgré cette différence de probabilité donnant l'écorce de pin comme une réalisation quasi-certaine, l'enveloppe de riz serait néanmoins à privilégier pour les tests. En effet, il est important de garder en tête que, comme introduit plus haut, l'une des limites des RB est son traitement des discrétisations : le comportement autour des valeurs limites peut donc être plus compliqué à prédire. Dans le cadre de l'écorce de pin, nous avons vu que son taux de cendres est trop bas (1.44 en moyenne), ce qui mitige déjà grandement les résultats prédits ; mais sa composition en lignine (27.33 en moyenne) le place tout juste au-dessus de la quantité de lignine recommandée par le modèle, rendant plus incertain son comportement. L'enveloppe de riz, au contraire, présente des compositions plutôt éloignées des limites des catégories de discrétisation. Il semble donc plus sûr de tester

p	0.99	p	0.82
Cendres	[6.7 ;24.7]	Cendres	[6.7 ;24.7]
Cellulose	[10.9 ; 25.6]	Cellulose	[25.6 ;33]
Lignine	[26.4 ; 49]	Lignine	[19.4 ; 26.4]
Exact	$\emptyset$	**Exact**	Enveloppe de Riz
Similaire	Ecorce de Pin	**Similaire**	$\emptyset$
S_{Pin}	5.26	S_{Riz}	0

TABLE 3 – Exemple de résultats correspondants aux critères d'acceptabilité définis, et leur propabilité p de réalisation. Lors de l'absence de correspondance exacte, un S-score a été calculé pour trouver le CL le plus proche de la cible.

	Cendres	Cellulose	Lignine
Ecorce de Pin	1.44	20.6	27.33
Enveloppe de Riz	14.5	31.9	25.7

TABLE 4 – Composition des deux nouvelles charges lignocellosiques candidates.

dans un premier temps ce matériel. Les compositions complètes des deux CLs sont présentées dans la Table 4.
En conclusion, si le choix des discrétisations établies a un sens pour le domaine, il introduit également des biais : l'appartenance d'une valeur à certaines catégories peut parfois être compliqué à distinguer, et certaines catégories semblent ainsi gonflées artificiellement par rapport à d'autres non représentées dans la BC. Cela souligne encore une fois l'importance de la représentativité d'une BC dans l'apprentissage automatique : une plus grande diversité de cas et d'exemples permettrait de limiter ces effets de bords.

5 Conclusion

Dans cet article, nous avons présenté POND, un workflow complet dédié à la réponse de questions expertes sur des bases de connaissance représentant des procédés de transformation modélisés par l'ontologie PO². Nous nous sommes focalisés ici sur la causalité, et les outils offerts par la découverte causale (comme l'ingénierie inverse), en présentant l'introduction de connaissances expertes à différents embranchements de la modélisation. Celle-ci se base sur deux points : l'établissement d'un vocabulaire commun standardisé à travers l'ontologie PO², et la formalisation de connaissances expertes ne pouvant pas être exprimées directement dans une BC car dépendantes du contexte. Nous avons ensuite illustré cette approche à travers une application concrète sur les emballages bio-composites. Grâce à l'ontologie, ce workflow permet à l'expert de facilement manier les connaissances expertes à intégrer d'une part, ainsi que l'ajout et la modification de celles-ci à la volée. Enfin, nous avons défini une formalisation de différentes contraintes expertes permettant de guider la lecture du RB appris afin d'éliciter les réponses les plus intéressantes du point de vue de l'utilisateur. Ainsi, à travers notre illustration, nous avons présenté plusieurs réponses possibles, et identifié de potentiels nouveaux matériaux à tester, encore non testés dans la base d'origine.

Comme dans toute analyse causale, il est important de considérer le contexte dans lequel l'apprentissage a lieu, et que nous avons détaillé en Section 2.3. Dans les travaux présentés ici, nous avons également supposé que l'expert consulté a une connaissance fiable du domaine, et où aucune contradiction n'est considérée (ce qui ne se serait pas forcément vérifié dans le cas d'un panel d'experts où des divergences peuvent se trouver). L'intégration de ces possibles dissensions et leur modélisation afin d'établir l'apprentissage d'un modèle optimal est une piste de recherche que nous souhaitons explorer.
De même, les recommandations établies par le RB et générées de façon automatique permettent d'établir une liste de règles établissant des scénarios plus ou moins crédibles. Par exemple, si l'on regarde la Table 2, il parait hautement improbable qu'une charge élevée (entre 21 et 50) permette d'obtenir une contrainte à la rupture améliorée (avec une probabilité quasi-nulle de 0.06). L'utilisation de ces règles pour évaluer la crédibilité de nouvelles informations ou la pertinence de la BC est une autre piste à étudier dans la poursuite de ces travaux.

Remerciements

Nous tenons à remercier Claire Meyer (Equipe PhyProDiv, INRAE IATE) pour nous avoir fourni les données pour la prospection de biomasses. Le travail présenté dans ce papier a été financé partiellement par l'Agence Nationale de Recherche dans le cadre des projets D2KAB (ANR-18-CE23-0017) et DataSusFood (ANR-19-DATA-0016).

Références

[1] Montassar Ben Messaoud, Philippe Leray, and Nahla Ben Amor. Semcado : A serendipitous strategy for learning causal bayesian networks using ontologies. *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 182–193, 2011.

[2] Giacomo Bucci, Valeriano Sandrucci, and Enrico Vicario. Ontologies and bayesian networks in medical diagnosis. *HICSS*, pages 1–8, 2011.

[3] Patrice Buche, Juliette Dibie-Barthelemy, Liliana L. Ibanescu, and Lydie Soler. Fuzzy Web Data Tables Integration Guided by a Termino-Ontological Resource. *IEEE Transactions on Knowledge and Data Engineering*, 25(4) :805–819, 2013.

[4] Federico Castelletti and Guido Consonni. Discovering causal structures in bayesian gaussian directed acyclic graph models. *Journal of the Royal Statistical Society Series A, Royal Statistical Society*, 183 :1727–1745, 2020.

[5] David Maxwell Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3(null) :507–554, mar 2003.

[6] Gregory F. Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4) :309–347, 1992.

[7] Grégoire David, Giovanna Croxatto Vega, Joshua Sohn, Anna Ekman Nilsson, Arnaud Hélias, Nathalie Gontard, and Helene Angellier-Coussy. Using life cycle assessment to quantify the environmental benefit of upcycling vine shoots as fillers in biocomposite packaging materials. *International Journal of Life Cycle Assessment*, 2020.

[8] C. P. De Campos and Q. Ji. Improving bayesian network parameter learning using constraints. In *ICPR*, pages 1–4, 2008.

[9] C.P. De Campos, Z. Zhi, and Q. Ji. Structure learning of bayesian networks using constraints. In *ICML*, pages 113–120, 2009.

[10] Juliette Dibie, Stéphane Dervaux, Estelle Doriot, Liliana Ibanescu, and Caroline Pénicaud. $[MS]^2O$ - A multi-scale and multi-step ontology for transformation processes : Application to micro-organisms. In *ICSS*, pages 163–176, 2016.

[11] Zhongli Ding, Yun Peng, and Rong Pan. *BayesOWL : Uncertainty Modeling in Semantic Web Ontologies*, pages 3–29. Springer Berlin Heidelberg, 2006.

[12] Lisa Ehrlinger and Wolfram Wöß. Towards a definition of knowledge graphs. In *SEMANTiCS (Posters, Demos, SuCCESS)*, 2016.

[13] Charlène Fabre, Patrice Buche, Xavier Rouau, and Claire Mayer-Laigle. Milling itineraries dataset for a collection of crop and wood by-products and granulometric properties of the resulting powders. *Data in Brief*, 33, 2020.

[14] Stefan Fenz. Exploiting experts' knowledge for structure learning of bayesian networks. *Data And Knowledge Engineering*, 73 :73 – 88, 2012.

[15] Nir Friedman, Lise Getoor, Daphne Koller, and Avi Pfeffer. Learning probabilistic relational models. In *IJCAI*, page 1300–1307. Morgan Kaufmann Publishers Inc., 1999.

[16] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10 :524, 2019.

[17] Alain Hauser and Peter Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, pages 926—-939, 2014.

[18] Liliana Ibanescu, Juliette Dibie, Stéphane Dervaux, Elisabeth Guichard, and Joe Raad. Po2- a process and observation ontology in food science. application to dairy gels. *Metadata and Semantics Research*, pages 155–165, 2016.

[19] Abdul-Wahid Mohammed. Knowledge-oriented semantics modelling towards uncertainty reasoning. *SpringerPlus*, 5, 2016.

[20] Melanie Munch, Patrice Buche, Cristina Manfredotti, Pierre-Henri Wuillemin, and Helene Angellier-Coussy. A process reverse engineering approach using Process and Observation Ontology and Probabilistic Relational Models : application to processing of biocomposites for food packaging. In *15th International Conference on Metadata and Semantics Research*, Madrid, Spain, November 2021.

[21] Melanie Munch, Juliette Dibie, Pierre-Henri Wuillemin, and Cristina E. Manfredotti. Towards interactive causal relation discovery driven by an ontology. In *FLAIRS*, pages 504–508, 2019.

[22] Melanie Munch, Juliette Dibie-Barthélemy, Pierre-Henri Wuillemin, and Cristina E. Manfredotti. Interactive causal discovery in knowledge graphs. In *Semex@ISWC 2019*, volume 2465 of *CEUR Workshop Proceedings*, pages 78–93. CEUR-WS.org, 2019.

[23] Melanie Munch, Pierre-Henri Wuillemin, Cristina Manfredotti, Juliette Dibie, and Stephane Dervaux. Learning probabilistic relational models using an ontology of transformation processes. In *OTM 2017 Conferences*, pages 198–215, 2017.

[24] Mélanie Munch, Patrice Buche, Stéphane Dervaux, and Hélène Angellier-Coussy. Itinerary Description for biocomposites from poly(3-hydroxybutyrate-co-3-hydroxyvalerate) and lignocellulosic fillers, 2021.

[25] Pekka Parviainen and Mikko Koivisto. Finding optimal bayesian networks using precedence constraints. *Journal of Machine Learning Research*, 14 :1387–1415, 2013.

[26] Judea Pearl. *Causality : Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009.

[27] Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2) :461 – 464, 1978.

[28] Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G Dimakis, and Sriram Vishwanath. Learning causal graphs with small interventions. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[29] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

[30] Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, and Mariano Fernández-López. The neon methodology for ontology engineering. In Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, Enrico Motta, and Aldo Gangemi, editors, *Ontology Engineering in a Networked World*, pages 9–34. Springer, 2012.

[31] Louis Verny, Nadir Sella, Séverine Affeldt, Param Singh, and Hervé Isambert. Learning causal networks with latent variables from multivariate information in genomic data. *PLOS Computational Biology*, 13, 2017.

[32] Shenyong Zhang, Yi Sun, Yun Peng, and Xiaopu Wang. Bayesowl : A prototype system for uncertainty in semantic web. *ICAI*, 2 :678–684, 2009.

Session 7 : Graphes de connaissances et temporalité

HHT : une ontologie modulaire pour représenter l'évolution des territoires en Histoire

L. Bourel[1], W. Charles[1], N. Hernandez [1 2], N. Aussenac-Gilles [1]

[1] IRIT- CNRS et Université de Toulouse prenom.nom@irit.fr
[2] Université Toulouse2 Jean Jaurès

31 mai 2022

Résumé

En histoire, modéliser la notion de territoire nécessite la prise en compte de ses différentes dimensions : hiérarchique, spatiale ou temporelle. L'ontologie HHT (Hierarchical Historical Territory) est proposée pour représenter l'état de fait de plusieurs découpages territoriaux hiérarchiques identifiés par la recherche historique. Définie au sein du projet ANR ObARDI qui étudie les territoires sous l'Ancien Régime, HHT repose sur différents modules permettant de représenter l'évolution des découpages territoriaux au cours du temps mais aussi les revendications ayant lieu sur ces découpages.

Mots-clés

Ontologie – Web sémantique – Humanités numériques – Objet spatio-temporel

Abstract

In digital humanities, modeling the concept of territory requires to take into account its different dimensions. The HHT ontology (Hierarchical Historical Territory) tries to represent the state of several hierarchical territorial divisions identified by historical research. In the ANR ObARDI project which studies the territories under the Ancien Régime, HHT comes in modular form to answer the specific need to represent historical evolutions of territorial divisions as well as claims on these divisions.

Keywords

Ontology – Semantic Web – digital humanities – spatio-temporel object

1 Introduction

La notion de territoire constitue un objet d'étude central à cheval entre l'histoire, la géographie et l'économétrie. Pour ces disciplines, il se caractérise par les éléments suivants :

- un pouvoir exercé par un acteur sur le territoire (**la domination**) ;
- l'espace dominé par ce contrôle territorial (**l'aire**) ;
- la connaissance des **limites** enserrant le territoire.

En conséquence, un territoire peut être un découpage administratif, un espace étatique, ou tout espace revendiqué par ses habitants [1]. Il s'agit toujours de l'appropriation d'un espace par un acteur.

Mais un territoire n'est pas réductible à sa seule délimitation spatiale ; la notion intègre également une dimension temporelle car le territoire évolue au cours du temps. Les paroisses de Paris en 1789 sont très différentes des communes - leurs équivalents actuels - tant par leur surface, leurs caractéristiques d'urbanisation, ou leur place dans les nomenclatures administratives. Pourtant, on peut établir un lien de filiation directe entre les paroisses et les communes. Par conséquent, il s'agit de parler d'un objet spatio-temporel dont l'évolution est déterminée par des activités humaines. Enfin, les territoires sont imbriqués dans des rapports de force, s'inscrivant dans une organisation territoriale complexe, parfois objet de revendications ou de conflits.

Dans cet article, nous nous intéressons à représenter cette notion de territoire à partir de ces différentes dimensions dans l'objectif d'y associer des informations et des connaissances qui pourront être de collectées, interrogées et analysées par des chercheurs en humanités numériques. Dans le domaine du web sémantique, les ontologies ont montré leur intérêt pour définir des vocabulaires partagés servant à décrire des entités d'un domaine en vue de les manipuler et de les lier. Nous présentons ici une ontologie permettant de représenter l'évolution des unités territoriales au cours du temps dans l'optique de construire un graphe de connaissances qui facilite aux chercheurs le partage de connaissances sur ces unités et l'étude de leurs évolutions.

Ce travail s'inscrit dans le cadre du projet ANR Obardi dont l'enjeu est d'améliorer la compréhension des dynamiques de pouvoir qui sous-tendent la construction de l'État moderne en France. En étudiant ses mécanismes de développement et sa manière de représenter un territoire politique, il s'agit de dépasser le méta-récit de la construction de l'État qui en freine encore sa compréhension [2] [5] [4]. La représentation du concept-clé qu'est le territoire est alors essentielle pour construire un graphe de connaissances servant de référentiel pour analyser les mécanismes de développement de l'Etat moderne.

L'ontologie HHT (*Historical Hierarchical Territory*) que

1. cf http://geoconfluences.ens-lyon.fr/glossaire/territoire
2. https://obardi.hypotheses.org/270/

nous proposons permet de représenter plusieurs découpages hiérarchiques simultanés du territoire et l'évolution de ces découpages territoriaux au cours du temps. Elle s'inspire fortement des ontologies TSN et TSN-Change [3] et présente l'originalité de permettre de représenter pour chaque unité territoriale sa propre temporalité d'évolution. Nous proposons également le module HHT-Claim afin de décrire les revendications et les conflits qui viennent s'opposer à l'état de fait historique décrit par HHT.
Dans la suite de l'article, la section 2 présentera la définition choisie des découpages en unités territoriales à travers le temps, et son impact en matière de modélisation. La section 3 exposera l'ontologie HHT construite à partir de cette réflexion, en la situant par rapport à l'état de l'art. La section 4 décrira le module HHT-Claim et la notion de revendication en histoire. La section 5 discutera de l'approche incrémentale envisagée pour le peuplement du graphe.

2 Représentation du territoire et de ses évolutions

Cette section détaillera l'état de l'art réalisé sur la représentation d'entités spatio-temporelles et les réflexions sur la représentation des objets inhérents à notre contexte.

2.1 Panorama de représentations d'entités spatio-temporelles et du territoire

L'ontologie OWL-time propose un vocabulaire standardisé par le W3C pour représenter le temps [10]. Elle permet de représenter des instants ou des intervalles temporels dans divers calendriers et d'exprimer des relations topologiques entre eux. Concernant l'espace, GeoSPARQL propose non seulement un vocabulaire mais aussi un mécanisme de raisonnement spatial [1]. Son vocabulaire permet de définir, à l'aide de coordonnées et d'une forme une zone dans l'espace. Néanmoins, dans notre cas, plus qu'une simple zone géographique, nous cherchons à définir un territoire.
En ce qui concerne la représentation du territoire, les ontologies sont nombreuses, et aucune ne semble faire autorité, chacune répondant à des besoins particuliers. Des ontologies d'applications sont spécifiquement dédiées à des découpages territoriaux administratifs. On peut citer ainsi l'ontologie proposée par l'INSEE qui rend compte du découpage administratif français actuel pour décrire un jeu de données sur le web des données liées [3]. Cependant, cette ontologie ne permet pas de représenter les évolutions progressives des territoires.
Or ceci est rendu possible par les ontologies TSN et TSN-Change [3] ou encore l'ontologie du projet SAMPO [11]. TSN permet de décrire un découpage pour l'ensemble d'un territoire à une période donnée tandis que TSN-change décrit les changements conduisant à passer d'une version de ce découpage territorial à une autre. TSN et TSN-Change reposent sur le postulat que le découpage de l'ensemble d'un territoire évolue à une date donnée. Sous l'Ancien Régime toutefois, les unités territoriales ont chacune leur propre temporalité d'évolution, ces évolutions étant décrites dans différentes sources. Dans TSN et TSN-Change, la trajectoire de vie des territoires est rythmée par des versions régulières de la nomenclature toute entière. Dans notre cas, chaque trajectoire de vie peut posséder ses propres références, ses propres sources. De ce point de vue, l'ontologie SAMPO intègre une approche bien plus dynamique [11], dont notre mécanisme de version s'inspirera. En effet, cette ontologie définit les états successifs d'un territoire, qu'elle rattache pour former un "ver spatio-temporel". A noter que le rattachement est ici réalisé sur la seule base du nom, quand la question de l'identité d'un territoire est autrement plus complexe (cf 2.4).
Enfin, CIDOC-CRM est une des ontologies de référence pour le patrimoine culturel [6]. Centrée autour de la notion d'évènement, CIDOC-CRM décrit plusieurs classes qui ont un intérêt pour notre cadre d'étude. Cependant, ni les niveaux ni la notion de territoire ne sont considérés.
Pour représenter les hiérarchies territoriales historiques, nous proposons une nouvelle ontologie réutilisant des concepts ou principes des vocabulaires cités plus haut.

2.2 Unités territoriales

Les historiens attribuent différentes caractéristiques aux territoires, qui peuvent toutes évoluer au cours du temps. Comme mentionné, un territoire est issu d'un rapport de force entre un acteur et un espace géographique. Les données démographiques, sociales ou les caractéristiques d'urbanisation d'un territoire sont des observations statistiques faites sur le territoire mais ne le caractérisant pas en lui-même. Un échange approfondi avec des historiens ainsi que l'étude des données usuelles associées aux territoires en histoire ont permis d'établir les propriétés caractérisant l'identité d'un territoire :

- Un nom
- Une géométrie (définissant sa délimitation spatiale)
- Son type ou sa catégorie hiérarchique (indiquant son rôle dans une hiérarchie donnée, tel que commune, département, région dans une hiérarchie administrative moderne ou paroisse, doyenné, archi-diaconné, évêché, archevêché dans une hiérarchie religieuse)
- Ses relations hiérarchiques avec d'autres territoires (comme, par exemple, la région dans laquelle se situe un département, ou les communes qui constituent un département). Ces relations hiérarchiques reflètent les rapports de force entre acteurs du territoire, i.e. le plus souvent entre les institutions en charge de la gouvernance de celui-ci. Elles sont cruciales car elles vont former le squelette hiérarchique des institutions de l'Ancien Régime, un des principaux objets d'étude dans le cadre d'ObARDI.

Une différence est à noter entre les notions de "territoire" et d'"unité territoriale". Cette dernière est une catégorie plus générale de la première, définie par les mêmes caractéristiques, à la différence près qu'elle ne considère pas l'influence dominante d'un individu ou groupe d'individus sur l'espace. L'unité territoriale correspond à un espace géographique pur, quand le territoire prend en compte sa signification dans une logique humaine. Cette distinction est

3. urlhttp ://rdf.insee.fr/

utile lorsque des catégories éthiques [4] sont à modéliser. Si par exemple, l'unité territoriale Midi-Pyrénées a toujours un sens géographique (on peut encore la dessiner sur une carte), le territoire Midi-Pyrénées n'existe plus, car l'institution en charge s'est dissoute lors de la formation des grandes régions.

2.3 Relations hiérarchiques

Une hiérarchie territoriale est considérée ici comme une classification des territoires. Elle repose sur un critère hiérarchique identifié par l'historien qui va servir de caractère discriminant permettant d'établir cette classification. Ces hiérarchies sont le reflet des relations hiérarchiques des territoires qui les constituent. De plus, cette relation hiérarchique de "domination" implique, en règle générale, une inclusion géographique pour les entités concernées. Après discussion avec les historiens, pour étudier les territoires de l'Ancien Régime, quatre dimensions de classification ont émergé : *administrative, religieuse, judiciaire, fiscale*. Ces quatre dimensions permettent d'établir quatre découpages hiérarchiques territoriaux différents du royaume de France, quatre filtres d'étude possibles qui se superposent les uns aux autres. Ils ne sont pas étanches entre eux puisque les unités territoriales peuvent posséder chacune une ou plusieurs dimensions. Chacune de ces dimensions possède ses propres niveaux hiérarchiques. Par exemple, le *découpage religieux* contient des archevêchés, des évêchés, des archidiaconés, des doyennés et des paroisses ecclésiastiques.

2.4 Identités et versions du territoire

Le territoire se trouve donc être un concept évoluant au cours du temps et impliqué dans des relations hiérarchiques avec d'autres territoires évoluant à des rythmes différents dans le temps.

Plusieurs options fondamentales existent pour approcher la modélisation d'objets évoluant dans le temps. Nous reprenons ici la dualité entre perdurance (propriété des entités qui ne changent pas dans le temps) et endurance (propriété des entités qui ont une durée déterminée, y compris instantanée) proposée par N. Guarino pour structurer les ontologies formelles [9]. La manière d'organiser une ontologie selon ces notions a conduit Grenon et Welty à définir deux types d'ontologies [8]. Pour rendre compte de la perdurance, les ontologies SPAN offrent une vision 4D des objets qui sont des "vers d'espace-temps" perdurant dans le temps. Construites pour représenter des entités endurantes, les ontologies SNAP, en revanche, donnent une vision tridimensionnelle d'un objet qui dure dans le temps, autrement dit, elles ne permettent pas de traduire son évolution. En reprenant les choix retenus dans les travaux sur TSN & TSN-Change [3], l'approche 4D-Fluent (une approche perdurantiste qui parle de *TimeSlice* pour représenter les versions du perdurant) semble la plus appropriée et correspond à notre définition du territoire établie avec les historiens du projet. Le territoire existera donc à travers toutes les versions de lui-même formant *la ligne de vie (ou trajectoire de vie)* de ce territoire. Mais chacune de ces versions possède des attributs et des caractéristiques qui lui sont propres.

Cependant, deux questions se posent :

— Quel est le critère d'identité diachronique qui permet d'établir l'identité du territoire ? [7]

— Quelles sont les caractéristiques d'une unité territoriale qui définissent la singularité de chacune de ses versions ?

Aucune des propriétés établies dans la section 2.2 (nom, géométrie, type, relations hiérarchiques) pour caractériser un territoire n'est essentielle pour son identité. Cependant, chacune de ces caractéristiques peut entraîner un changement [7] qui engendre une nouvelle version. Ces changements sont non-disruptifs, c'est-à-dire ne modifient pas l'identité du perdurant représentant la ligne de vie d'un territoire. Ces quatre caractéristiques définissent donc la singularité de chacune des versions considérées, répondant à notre deuxième question.

Reste à déterminer le critère permettant de qualifier un *changement disruptif*. Ce critère ne doit rester qu'un guide. En effet, dans une démarche de recherche historique, seul l'historien peut, grâce à l'analyse des sources, trancher sur la véracité ou non d'une connaissance historique. Néanmoins, il est utile d'expliciter le critère diachronique d'identité que l'on veut justement mettre à l'épreuve de l'analyse des historiens. En reprenant celui développé par Garbazc [7], nous énonçons ainsi le critère d'identité local : *Changement disruptif il y a si et seulement si le nom s'en retrouve modifié en même temps qu'une autre caractéristique du territoire (catégorie, géométrie, relation hiérarchique). Tout autre changement dans lequel serait impliqué un territoire sera non disruptif.*

Néanmoins, ce critère n'est pas global. On peut attester de la même identité entre deux versions de territoire si et seulement si l'état des connaissances permet de déterminer l'ensemble des changements non disruptifs qui ont modifié les qualités de ce territoire pour le faire passer de la première à la seconde version.

3 Ontologie HHT

Nommée HHT pour *Hierarchical Historical Territory*, notre ontologie s'inspire très fortement des ontologies TSN et TSN-Change mais s'en éloigne sur des points fondamentaux. Elle est disponible à l'adresse `https://www.irit.fr/recherches/MELODI/ontologies/ObARDI/`.

3.1 Spécification de l'ontologie

La première étape dans le développement d'une ontologie, selon la méthodologie NEON [12], est d'identifier les exigences de celle-ci. En premier lieu, une discussion poussée avec les experts du domaine, ici les historiens, a permis de cerner plus finement l'objet d'étude et ses caractéristiques qui ont été détaillés dans les sections précédentes.

Les objectifs auxquels HHT tente de répondre sont alors de modéliser les éléments suivants :

— les unités territoriales ;

— leurs niveaux hiérarchiques ;

4. correspondant à un filtre du point de vue de l'observateur et non pas de la réalité historique de l'époque

— différents critères de classification hiérarchique ;
— les changements subis par ces unités et ces niveaux ;
— l'évolution des connaissances des historiens.

3.2 Hiérarchies territoriales

Cette première partie de l'ontologie vise à représenter n'importe quelle hiérarchie territoriale historique. Ainsi, les classes définies sont purement génériques, et peuvent être appliquées à tout contexte. Les éléments relatifs aux spécificités de l'objet d'étude d'ObARDI (critères hiérarchiques, niveaux hiérarchiques, etc.) sont quant à eux représentés à l'aide d'instances. Ainsi, l'ontologie HHT repose uniquement sur des concepts transposables à diverses périodes :

`hht:Unit` Sous-classe de `hht:Area` (une simple zone géographique), représentant un espace géographique appartenant à une hiérarchie. Ce concept est défini par sa dimension spatiale et par sa dimension hiérarchique.

`hht:historicalTerritory` Sous-classe de `hht:Unit` représentant une portion de l'espace géographique réclamé ou occupé par une personne, un groupe de personnes ou une institution qui en définit elle-même les frontières.

`hht:HierarchicalCriterion` Critère hiérarchique, caractère discriminant qui permet de définir un découpage hiérarchique de l'espace en différents niveaux hiérarchiques (`hht:Level`) permettant de classifier des `hht:Unit`.

Chacun de ces concepts permet de décrire l'espace. Pour prendre en compte leur évolution dans le temps, nous définissons 3 nouveaux concepts, versions des concepts précédents, possédant chacun leur propre période de validité (via la propriété `hht:validityPeriod`). Un mécanisme de versionnage associé est détaillé section 5.

`hht:UnitVersion` Version d'une `hht:Unit` sur une période de validité donnée. Elle possède des unités supérieures auxquelles elle est liée par la propriété `hht:hasSuperUnit` et des unités inférieures auxquelles elle est liée par la propriété `hht:hasSubUnit`.

`hht:HistoricalTerritoryVersion` Version d'un `hht:HistoricalTerritory` sur une période de validité donnée.

`hht:LevelVersion` Version d'un niveau hiérarchique sur une période de validité donnée. Un niveau possède un niveau supérieur (`hht:hasSuperLevel`) et un niveau inférieur (`hht:hasSubLevel`). Un niveau hiérarchique possède des `htt:UnitVersion` par la propriété `htt:hasMember`.

La figure 1 représente les concepts ainsi que les relations définies dans le module.

3.3 Evolutions territoriales

Cette deuxième partie de l'ontologie vise à représenter les changements ayant mené à la création d'une nouvelle version d'unité territoriale. Trois grands types de changements sont considérés :

— `hht:FeatureChange` : représentant la modification d'une simple caractéristique entre deux versions d'une unité territoriale (expansion, contraction, changement de nom, apparition, changement d'unité supérieure, etc.).
— `hht:CompositeChange` : changement composite, c'est-à-dire un évènement regroupant ou enchaînant plusieurs changements simples.
— `hht:UpdateKnowledge` : Cet évènement est différent des deux autres dans le sens où il ne décrit pas un évènement historique mais la mise à jour du graphe de connaissances. Il permet de retracer ainsi l'ensemble des versions par lesquelles est passée l'information sur une unité territoriale pour arriver à l'état actuel des connaissances. Seule la dernière version est considérée valide tandis que les autres sont vues comme des connaissances obsolètes.

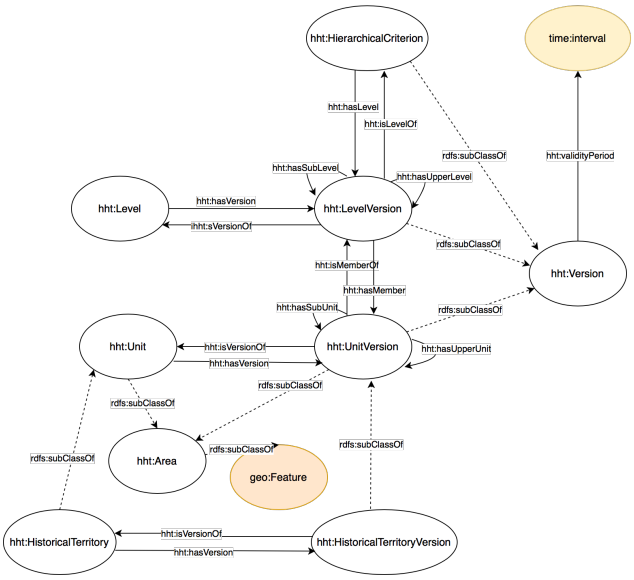


FIGURE 1 – Schéma de l'ontologie HHT

Chacun de ces concepts est une sous-classe du concept E5 Évènement [5] de CIDOC-CRM .

4 HHT-Claim

L'ontologie HHT peut être étendue par divers modules pour répondre à des besoins spécifiques, comme montré dans cet article avec les notions de revendication et de conflit territorial.

4.1 Revendication

La brique de base d'un conflit, tel que défini dans ObARDI, est une revendication d'une unité territoriale par un acteur (`hht:Claim`).
La hiérarchie modélisée par HHT représente ce que les historiens nomment *l'état de fait*, mais L'Ancien Régime regorge de volontés plus ou moins locales d'altérer celui-ci. Modéliser ces conflits et ces relations hiérarchiques conflictuelles est de grande importance pour aider à l'analyse de

5. https ://cidoc-crm.org/Entity/e5-event/version-6.2

l'historien.
Une revendication peut être définie comme le souhait d'un acteur d'altérer une ou plusieurs relations hiérarchiques qu'entretient un territoire avec un autre. Toute revendication est caractérisée par les propriétés `hht:upperTerritory` et `hht:subTerritory` qui indiquent respectivement le territoire supérieur et inférieur impliqué dans la revendication. Une propriété `hht:validityPeriodOfClaim` indique la période durant laquelle cette revendication est valide. Enfin, une propriété `hht:makeAClaim` permet de rattacher un acteur à une revendication. On distingue dans HHT-Claim trois types de revendications :

— `DeclarationUnder` : Un acteur du territoire inférieur souhaiterait se placer sous une juridiction plus avantageuse pour lui.
— `ClaimTo` : Un acteur du territoire supérieur cherche à placer le territoire inférieur sous sa juridiction.
— `AutonomyRequest` : Une revendication sans territoire supérieur, car cherchant justement la création d'un territoire supérieur pour gouverner le territoire inférieur.

4.2 Zone de conflit

Émerge également une notion liée au conflit : la zone dans laquelle celui-ci se produit. En effet, les revendications rattachées à plusieurs territoires adjacents peuvent dans certains cas être similaires par leur nature ou leur cause. Pour représenter ces similarités, HHT-Claim comporte la notion de zone de conflit, définie comme une agrégation de revendications. Cette notion de zone de conflit soulève plusieurs interrogations. Tout d'abord, une interrogation de temporalité. Si les revendications possèdent une période de validité, il est nécessaire d'en attribuer une à une zone de conflit. Afin d'obtenir une représentation fine des zones de conflit, nous serons donc amenés à considérer dans une version ultérieure du module HHT-Claim, l'ajout de versions des zones de conflits, comme c'est déjà le cas pour les territoires. Autre interrogation : la caractérisation intrinsèque de ces zones de conflits. L'enjeu ici serait la détermination d'un critère permettant de définir sémantiquement et automatiquement les limites d'une zone de conflit d'après les revendications considérées.

5 Construction incrémentale du graphe de connaissances

Si l'ontologie HHT est une ontologie générique, son utilisation dans le cadre projet ObARDI est toutefois à noter. Le graphe de connaissances produit dans ce contexte est en effet construit de manière continue et incrémentale au fur et à mesure que de nouvelles connaissances sont produites ou rectifiées par des chercheurs lorsqu'ils analysent des sources. L'état de chaque ligne de vie des territoires doit alors se modifier en conséquence, selon les nouvelles informations apportées et intégrées au graphe de connaissances. Non seulement ce graphe cherche à modéliser l'état actuel

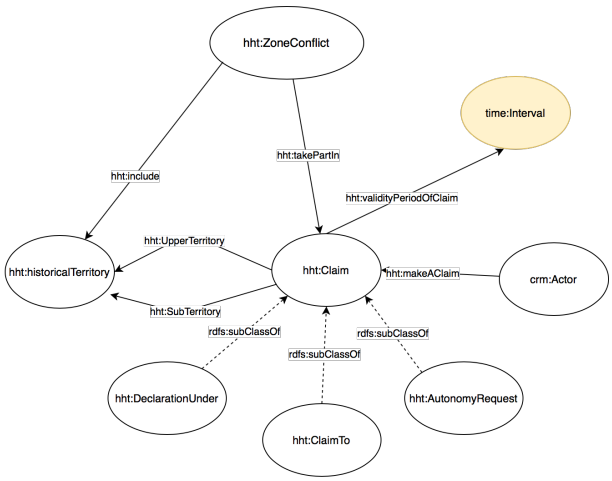


FIGURE 2 – Schéma du module HHT-Claim

des connaissances, mais il conserve également chacun des états passés de connaissance. Plus exactement, lorsqu'une nouvelle connaissance sur une unité territoriale est ajoutée et vient contredire/compléter des informations déjà disponibles sur cette période, la version porteuse des précédentes informations sera conservée.

5.1 Construction du graphe

Dans le cadre du projet ObARDI, la majorité des données sont fournies par des historiens, des géographes et des économistes. Aussi faut-il leur fournir un outil leur permettant d'ajouter facilement des données. Dans cette optique, un portail sémantique a été développé, permettant aux utilisateurs d'ajouter un territoire/une version d'un territoire de manière individuelle, ou de réaliser une importation globale à partir de fichier .csv. Le graphe de connaissances peut ensuite être interrogé à l'aide d'une fonctionnalité de recherche permettant de rechercher des territoires selon plusieurs critères (nom, critère, niveau, date, etc.)
Le développement de ce portail sémantique met toutefois en lumière un certain nombre de problématiques dues à l'évolution incrémentale du graphe de connaissance. En particulier, les raisonnement réalisés lors de l'ajout de connaissances le sont sur des données incomplètes. Ainsi, ces raisonnements doivent être réalisés à l'ajout de chaque unité territoriale, ce qui alourdit considérablement le traitement, notamment lors d'un enrichissement via un fichier.

5.2 Version de connaissance obsolète

A chaque insertion d'une entité sur un intervalle temporel séquent à celui d'une version existante, ladite version devient obsolète dans le graphe de connaissances, au sens où l'on possède une nouvelle information considérée comme plus véridique. Lorsque cela arrive, au lieu de la supprimer du graphe, on modifie simplement la valeur booléenne de la propriété *hht :isDeprecated* pour indiquer de ne plus la prendre en compte. De plus, on note la liste des modifications apportées à une unité par des changements *UpdateKnowledge*. Garder l'ensemble des connaissances obsolètes

alourdit le graphe mais permettra d'analyser l'évolution des connaissances durant le temps du projet ObARDI.

6 Conclusion

L'ontologie HHT s'attache à représenter des hiérarchies territoriales flexibles et adaptées à la recherche en histoire. Elle ne se concentre pas que sur l'aspect géométrique et hiérarchique de ces territoires, car elle vise à les représenter, indépendamment de l'existence d'une source de données décrivant l'entièreté du territoire.
HHT se place dans une perspective historique, s'inscrivant dans les humanités numériques, tandis que TSN est à portée statistique pour proposer une norme commune à tous les découpages territoriaux actuels. Dans cet objectif, HHT cherche également à modéliser les états passés des connaissances historiques. Le module HHT-Claim étend l'ontologie HHT pour représenter les revendications et les conflits ayant lieu sur ces territoires. Ce module permettra d'analyser les conflits et les relations entre acteurs, et aurait donc à terme une place centrale dans l'utilisation de l'ontologie HHT.
D'autres modules peuvent venir également s'ajouter à l'ontologie HHT pour préciser des informations connexes aux territoires décrits par HHT, comme la gestion des sources historiques. Dans ce cas notamment, en se basant sur des approches telles que celle adoptée dans SyMOGiH [2], il serait possible de garder chaque interprétation produite afin de calculer l'état de connaissance le plus probable. Dans cette optique, il sera donc nécessaire de conserver les états précédents du graphe de connaissance tout en exposant un état de celui-ci représentatif de l'état des connaissances. cette dimension projet pousse à envisager un mécanisme de versionnage du graphe de connaissance qui n'est pas sans rappeler celui offert par Git.
Enfin, un algorithme de raisonnement permettant de qualifier les changements liant deux versions consécutives devra être réalisé, en prenant en compte les difficultés liées à l'ajout incrémental de connaissances.
De nombreuses pistes restent donc encore à explorer pour affiner et étendre ce modèle.

7 Remerciements

Ce travail se déroule dans le cadre du projet ObARDI, financé par l'Agence Nationale de la Recherche de janvier 2021 à janvier 2025 dans le cadre de l'Appel à projets générique 2020. Il se situe dans le CE38 (Révolution numérique : rapports au savoir et à la culture), axe interdisciplinaire liant sciences du numérique et SHS.

Références

[1] Robert BATTLE et Dave KOLAS. "Geosparql : enabling a geospatial semantic web". In : *Semantic Web Journal* 3.4 (2011), p. 355-370.

[2] Francesco BERETTA. *L'interopérabilité des données historiques et la question du modèle : l'ontologie du projet SyMoGIH*. Presses universitaires de Paris Nanterre, 2017. ISBN : 978-2-84016-268-1. URL : `https://halshs.archives-ouvertes.fr/halshs-01559816` (visité le 11/05/2021).

[3] Camille BERNARD. "Immersing evolving geographic divisions in the semantic Web". Thèse de doct. Université Grenoble Alpes, 2019.

[4] Willem Pieter BLOCKMANS, André HOLENSTEIN et Jon MATHIEU. *Empowering interactions : political cultures and the emergence of the state in Europe, 1300-1900*. Ashgate Publishing, Ltd., 2009.

[5] Wim BLOCKMANS. *The origins of the modern state in Europe : 13th to 18th centuries*. Clarendon Press, 1995.

[6] George BRUSEKER, Nicola CARBONI et Anaıs GUILLEM. "Cultural heritage data management : the role of formal ontology and CIDOC CRM". In : *Heritage and Archaeology in the Digital Age* (2017), p. 93-131.

[7] Pawel GARBACZ, Bogumił SZADY et Agnieszka ŁAWRYNOWICZ. "Identity of historical localities in information systems". In : *Applied Ontology* 16.1 (2021), p. 55-86.

[8] Pierre GRENON et Barry SMITH. "SNAP and SPAN : Towards Dynamic Spatial Ontology". In : *Spatial Cognition & Computation* 4 (2004), p. 104-69.

[9] Nicola GUARINO. "Some Ontological Principles for Designing Upper Level Lexical Resources". In : *First International Conference on Language Resources and Evaluation, Granada, Spain*. T. 1. Juin 1998, p. 527-534.

[10] Jerry R HOBBS et Feng PAN. "Time ontology in OWL". In : *W3C working draft* 27 (2006), p. 133.

[11] Eero HYVÖNEN et al. "Representing and utilizing changing historical places as an ontology time series". In : *Geospatial Semantics and the Semantic Web*. Springer, 2011, p. 1-25.

[12] Mari Carmen SUÁREZ-FIGUEROA, Asunción GÓMEZ-PÉREZ et Mariano FERNÁNDEZ-LÓPEZ. "The NeOn methodology for ontology engineering". In : *Ontology engineering in a networked world*. Springer, 2012, p. 9-34.

Traitement des données temporelles certaines et incertaines en OWL 2 : Approche basée sur la théorie des probabilités

Nassira Achich[1,2], Fatma Ghorbel[1,2], Fayçal Hamdi[2], Elisabeth Métais[2], Faïez Gargouri[1]

[1] Laboratoire MIRACL, Université de Sfax, Tunisie
[2] Laboratoire CEDRIC, Conservatoire National des Arts et Métiers, Paris, France

Résumé

Les données temporelles fournies par les patients atteints Alzheimer sont sujettes à l'incertitude. De nombreuses approches ont été proposées pour traiter des données temporelles certaines, mais non pas celles qui sont incertaines. Cet article propose une approche pour représenter et raisonner sur des intervalles et des points de temps quantitatifs certains et incertains et les relations qualitatives entre eux. Elle inclut trois volets : (1) une extension de l'approche 4D-fluents avec de nouvelles composantes ontologiques pour représenter des données temporelles certaines et incertaines. (2) une extension de l'algèbre des intervalles d'Allen pour raisonner sur des intervalles de temps certains et incertains. Une adaptation de relations cette algèbre pour relier un intervalle de temps et un point de temps, et deux points de temps. (3) Enfin une ontologie qui intègre toutes ces extensions. Un prototype a été implémenté et intégré dans une prothèse de mémoire pour les patients atteints de la maladie d'Alzheimer afin de gérer des entrées de données incertaines.

Mots-clés

Approche 4D-Fluents, Algèbre d'intervalle d'Allen, Réseaux bayésiens, OWL 2, Raisonnement temporel, Représentation temporelle, Données temporelles incertaines.

Abstract

Temporal data given by Alzheimer's patients are subject to uncertainty. Many approaches have been proposed to deal with certain temporal data and neglect uncertain ones. This paper proposes an approach to representing and reasoning about certain and uncertain quantitative intervals and time points and the qualitative relations between them. It includes three contributions. (1) We extend the 4D-fluent approach with new ontological components to represent certain and uncertain temporal data. (2) We extend Allen's interval algebra to reason about certain and uncertain time intervals. We adapt these relations to relate a time interval and a time point, and two time points. (3) We propose an ontology based on our extensions. A prototype is implemented and integrated into an ontology-based memory prosthesis for patients with Alzheimer's disease to handle uncertain data inputs.

Keywords

4D-Fluents Approach, Allen's Interval Algebra, Bayesian Network, OWL 2, Temporal Reasoning, Temporal Representation, Uncertain Temporal Data.

1 Introduction

Les données temporelles sont sujettes à l'incertitude. En effet, ce genre d'imperfection pourrait être fréquent surtout dans le cadre d'une application comme CAPTAIN MEMO [31], qui est une prothèse de mémoire pour les patients d'Alzheimer. Cette prothèse offre un ensemble de services tels que l'aide aux utilisateurs pour se souvenir de leurs proches. Elle est basée sur une ontologie OWL 2, appelée « *PersonLink*[1] » [23]. Cette dernière permet de modéliser et de raisonner sur des relations interpersonnelles (e.g. parent, voisin) et de décrire des personnes. Un patient Alzheimer peut saisir dans CAPTAIN MEMO des données incertaines telles que « *Je pense que c'était de 2013 à 2020* », « *Elle a quitté le pays peut-être à 11 heures* » et « *Je pense qu'elle s'est mariée avant l'obtention du diplôme universitaire* ».

Dans le domaine du Web sémantique, plusieurs approches ont été proposées pour représenter et raisonner sur des données temporelles certaines. Cependant, la plupart d'entre elles ne traitent que des intervalles de temps et les relations qualitatives entre eux, c'est-à-dire qu'ils ne sont pas destinés à gérer des points de temps et des relations qualitatives entre un intervalle de temps et un point de temps ou deux points de temps. Par ailleurs, à notre connaissance, il n'existe pas d'approches dédiées au traitement des données temporelles incertaines dans OWL 2.

Dans cet article, nous présentons notre approche proposée dans [1], basée sur OWL 2, pour représenter et raisonner sur des données temporelles certaines et incertaines en termes de relations qualitatives et quantitatives, avec une vue certaine. Cette approche inclut trois volets. Le premier concerne la représentation des données temporelles certaines et incertaines

[1] http://cedric.cnam.fr/isid/ontologies/PersonLink.owl#

en OWL 2. Pour cela nous avons étendu l'approche 4D-fluents [40], qui ne couvre que les intervalles de temps certains dans OWL, avec de nouvelles composantes ontologiques certaines pour représenter des données temporelles quantitatives certaines et incertaines, et des relations temporelles qualitatives entre les intervalles de temps et les points. Nous avons utilisé les réseaux bayésiens pour calculer les mesures de certitudes. Le deuxième volet concerne le raisonnement sur les données temporelles certaines et incertaines en étendant l'algèbre d'intervalle d'Allen [3]. Cette algèbre ne prend en compte que les relations temporelles entre intervalles de temps certains. Nous l'étendons pour gérer les relations temporelles entre des intervalles de temps incertains et des points temporels certains et incertains. Notre extension préserve des propriétés importantes de l'algèbre d'origine. Nous adaptons les relations entre intervalles pour proposer des relations temporelles entre un intervalle de temps et un point de temps, et deux points de temps. Toutes ces relations sont utilisées par la suite pour le raisonnement temporel via des tables de transitivité. Enfin le dernier volet est la proposition d'une ontologie OWL 2 appelée « *UncertTimeOnto* » qui peut être intégrée dans d'autres ontologies pour traiter des données temporelles certaines et incertaines. Elle a été implémentée sur la base des extensions proposées. Les inférences sont effectuées à l'aide de règles SWRL intégrées dans l'ontologie.

Le reste de cet article est organisé comme suit. Les travaux proches sont passés en revue dans la section 2. La section 3 présente notre extension de l'approche 4D-fluents pour représenter les données temporelles incertaines dans OWL 2. La section 4 présente notre extension de l'algèbre d'Allen pour raisonner sur les données temporelles incertaines. La section 5 présente notre ontologie « UncertTimeOnto » basée sur nos extensions. La section 6 introduit les expérimentations de notre approche dans le cadre de CAPTAIN MEMO. Dans la dernière section, nous terminons par une conclusion.

2 Travaux proches

Les données temporelles peuvent être certaines ou incertaines. Elles sont également caractérisées en termes quantitatifs ou qualitatifs. Des données temporelles quantitatives certaines signifient des intervalles de temps et des points de temps certains. Des données temporelles quantitatives incertaines signifient des intervalles de temps et des points de temps incertains. Les intervalles de temps incertains sont des intervalles de temps normaux caractérisés par des limites de début et/ou de fin incertaines (par exemple, « *Peut-être de 2015 à 2018* »). Les points de temps incertains sont des points de temps qui sont définis de manière incertaine (par exemple, « *Je ne suis pas sûr si c'était en 2010* »). Les relations temporelles qualitatives lient deux intervalles de temps (Intervalle-Intervalle), un intervalle de temps et un point de temps (Intervalle-Point et Point-Intervalle) ou deux points de temps (Point-Point) ; où les intervalles de temps et les points peuvent être certains ou incertains. Les données temporelles qualitatives peuvent également être certaines ou incertaines et peuvent être déduites de données quantitatives.

Dans cette section, nous introduisons les travaux connexes relatifs à la représentation et au raisonnement sur les données temporelles dans le web sémantique.

2.1 Représentation des données temporelles dans le web sémantique

La représentation des données temporelles sous forme d'ontologie est un besoin crucial. Cependant, les langages ontologiques tels que OWL fournissent un support minimal, car ils sont tous basés sur des relations binaires qui relient simplement deux instances. Cela explique l'émergence de nombreuses approches pour représenter et raisonner sur les données temporelles dans le domaine du web sémantique.

Nous avons identifié dans la littérature plusieurs approches permettant la modélisation de la dimension temporelle dans le domaine du web sémantique : les logiques de description temporelles [5], le versioning [28], la réification [10], les relations N-aires [33] et l'approche 4D-fluents. Nous avons classé les approches en deux catégories : (i) les approches qui étendent la syntaxe OWL ou RDF en définissant de nouveaux opérateurs pour incorporer les données temporelles, et (ii) les approches qui sont implémentées directement en utilisant OWL ou RDF pour représenter les données temporelles sans étendre leurs syntaxes. La première catégorie comprend la logique de description temporelle, les « *concrete domains* » et le RDF temporel [21].

La logique de description temporelle étend les logiques de description standard avec de nouvelles sémantiques temporelles telles que « *until* ». Cette approche conserve la décidabilité et ne conduit pas à une redondance au niveau de la représentation des données. Cependant, elle nécessite l'extension de OWL ou RDF, qui est une tâche fastidieuse. Les domaines concrets [30] est aussi une approche qui nécessite l'introduction de types de données et d'opérateurs supplémentaires dans OWL. OWL-MeT [11] et TL-OWL [27] sont des implémentations de cette approche. Le RDF temporel, qui appartient également à la première catégorie, n'utilise que des triplets RDF. Il n'a pas toute l'expressivité du langage OWL et ne permet pas d'exprimer des relations qualitatives. Dans [26], les auteurs présentent un cadre complet pour incorporer le raisonnement temporel dans RDF. [29] introduisent un modèle de données de contraintes pour représenter les données temporelles, qui étend RDF, nommé stRDF. Ils étendent SPARQL pour interroger stRDF.

Le versioning, la réification, « *N-ary relations* », « 4D-Fluents » et les graphes nommés [39] sont des approches qui appartiennent à la deuxième catégorie. Le versioning est décrit comme la capacité de gérer les changements dans l'ontologie en créant différentes versions de l'ontologie. [41] proposent une approche pour la gestion des versions de schéma dans OWL. Son principal problème est que toutes les versions sont indépendantes les unes des autres ce qui nécessite des recherches exhaustives dans chacune d'elles. La réification est une technique de représentation des relations N-aires lorsque seules les relations binaires sont autorisées. Un nouvel objet est créé chaque fois qu'une relation temporelle doit être représentée. « *N-ary relations* » est approche qui propose de représenter une relation N-aire comme deux propriétés liées chacune à un nouvel objet tout en maintenant la sémantique des propriétés. Cette approche souffre d'un problème de redondance des données. [35] proposent un plug-in pour

PROTÉGÉ, nommé CHRONOS, qui utilise cette approche pour gérer les données temporelles. Une solution aux problèmes de redondance, dont souffre « N-ary relations », a été proposé par l'approche 4D-fluents qui représente les intervalles de temps et leur évolution dans OWL. Dans cette approche les changements ne se produisent que dans les parties temporelles et les concepts variant dans le temps sont représentés comme des objets à 4 dimensions avec la 4ème dimension étant les données temporelles. Plusieurs travaux ont utilisé l'approche 4D-fluents. La plupart d'entre eux, qui sont [34], [7], [22], [4], [8] et [24], l'ont utilisé dans le contexte de données temporelles considérées comme précises et certaines, ce qui ne correspondent pas à notre contexte. D'autres l'ont étendue pour représenter des données temporelles imparfaites, en particulier des données temporelles imprécises comme [16], qui ne correspondent pas non plus à notre contexte. A notre connaissance, il n'existe aucune approche pour représenter des données temporelles incertaines dans OWL 2. L'approche des graphes nommés représente chaque intervalle de temps par exactement un graphe nommé, où tous les triplets appartenant partagent la même période de validité.

Toutes les approches passées en revue ne traitent que les données temporelles certaines et négligent celles qui sont incertaines. Elles ne sont pas destinées à gérer des points de temps et des relations temporelles qualitatives entre un intervalle de temps et un point de temps ou même deux points de temps. Un critère de base dans le choix de notre approche est qu'elle doit s'appuyer sur des constructeurs déjà définis dans OWL. Par conséquent, nous excluons les approches « *Temporal Description Logic* », « *Concrete Domain* » et « *Temporal RDF* ». Nous excluons également l'approche « *Named Graphs* », car elle ne prend pas en charge OWL et n'est pas une solution conforme aux standards W3C.

Notre choix s'est porté sur l'approche 4D-fluents que nous avons étendue pour représenter des données temporelles quantitatives incertaines et les relations temporelles qualitatives associées. Comparée aux approches de réification, de « N-ary relation » et de versioning, l'approche 4D-fluents minimise le problème de redondance des données, car les changements ne se produisent que sur les parties temporelles et maintiennent la partie statique inchangée.

2.2 Raisonnement sur les données temporelles dans le web sémantique : L'Algèbre d'intervalle d'Allen

13 relations temporelles qualitatives entre des intervalles de temps classiques sont proposées par Allen. Elles sont définies en termes d'ordre des bornes de début et de fin des intervalles correspondants (voir tableau 1). Une particularité de l'algèbre d'Allen est que l'on peut déduire de nouvelles relations à travers la composition d'autres relations (par exemple, « Before (A, B) » et « Equals (B,C) » donne « Before(A, C) »). Cependant cette algèbre ne traite pas des intervalles de temps incertains. De plus, elle ne permet pas de lier un point de temps et un intervalle de temps, ni deux points de temps.

Plusieurs approches ont été proposées pour étendre cette algèbre. Certaines proposent des relations temporelles entre des intervalles de temps précis comme [13], [45], [46] et [47]. Dans [20], les auteurs proposent des relations temporelles floues vues comme des ensembles flous de relations d'Allen comme la relation « *Fuzz-Meets* » qui concerne la relation d'Allen « *Meets* ». Dans [36] les auteurs proposent une approche basée sur la théorie des possibilités pour modéliser des relations d'intervalles temporels incertains en attribuant un degré de préférence à chaque relation d'Allen de base. Cette approche est proposée dans le cadre d'un réseau probabiliste d'intervalles temporels ; où les nœuds sont des intervalles de temps et les arêtes sont des relations d'intervalle incertaines. [6] propose une extension floue de l'algèbre d'Allen. Dans ce travail, les auteurs associent un degré de préférence à chaque relation temporelle.

Certaines approches proposent des relations temporelles entre des intervalles de temps imprécis comme [32], [37] et [14]. Cependant, ces extensions sont basées sur des théories liées aux données imparfaites et ne peuvent être supportées dans le cadre de certaines ontologies. Par exemple, les approches proposées dans [33] et [14], sont basées sur la théorie des ensembles flous. De plus, la plupart de ces extensions ne conservent pas toutes les propriétés de l'algèbre d'Allen originale. Par exemple, dans [33], la relation « *Equals* » n'est pas réflexive. Les compositions des relations résultantes ne sont pas étudiées par les auteurs. [14] généralisent les relations d'Allen pour les rendre applicables à des intervalles de temps imprécis de manière conjonctive et disjonctive. Cependant, les auteurs ne proposent pas de tableau de composition des relations proposées. La représentation et le raisonnement sur des données temporelles existent également dans d'autres domaines tels que les bases de données, les images de télédétection et la récupération d'informations existent [48] et [49].

Dans nos travaux précédents [2], nous avons étendu l'algèbre d'intervalle d'Allen pour raisonner sur des dates et des horloges précises et imprécises dans le cadre d'une ontologie dite « certaine ».

Tableau 1. Relations d'Allen entre deux intervalles de temps précis

Relation (I, J)	Relations entre les bornes	Illustration	Inverse
Before	$I^+ < J^-$		After
Meets	$I^+ = J^-$		Met-by
Overlaps	$(I^- < J^-) \wedge (I^+ > J^-) \wedge (I^+ < J^+)$		Overlapped-by
Starts	$(I^- = J^-) \wedge (I^+ < J^+)$		Started-by
During	$(I^- > J^-) \wedge (I^+ < J^+)$		Contains
Ends	$(I^- > J^-) \wedge (I^+ = J^+)$		Ended-by
Equals	$(I^- = J^-) \wedge (I^+ = J^+)$		Equals

Les extensions indiquées sont basées sur des théories liées aux données imparfaites. Par conséquent, les relations résultantes ne peuvent pas être traitées dans le contexte d'une ontologie « certaine ». De plus, la plupart de ces approches ne conservent pas les propriétés de l'algèbre d'origine et n'étudient pas la composition des relations résultantes.

3 Représentation des données temporelles certaines et incertaines dans OWL 2

Nous avons étendu l'approche 4D-fluents, avec des composantes ontologiques « certaines », pour représenter en OWL 2 les données temporelles quantitatives « certaines » et « incertaines » ainsi que les relations temporelles qualitatives associées.

3.1 Données temporelles quantitatives

Nous avons étendu l'approche 4D-fluents pour représenter *(i)* des intervalles de temps incertains et *(ii)* des points de temps certains et incertains. La Figure 1 illustre notre extension.

3.1.1 Intervalles de temps incertains

Soit $A = [A^{-}_{a-}, A^{+}{}_{a+}]$ un intervalle de temps incertain. A^{-} et A^{+} sont respectivement les bornes supérieures et inférieures ; et a^{-} et a^{+} sont, respectivement, les degrés de certitude associés. Nous étendons l'approche 4D-fluents pour représenter des intervalles de temps incertains en introduisant deux propriétés de type de données (*Datatype Property*) « certaines » nommées « *HasBeginnigCertainty* » et « *HasEndCertainty* ». Ces deux propriétés sont associées à la classe prédéfinie « *TimeInterval* ». « *HasBeginnigCertainty* » et « *HasEndCertainty* » représentent, respectivement, les degrés de certitude associés à la borne supérieure a^{-} et à la borne inférieure a^{+}. Par exemple, si nous avons les informations suivantes données par un patient atteint de la maladie d'Alzheimer : « *Je pense que Marie a enseigné à l'Université de Paris Sorbonne de 2001 à 2010* ». Dans ce cas « $[2001_{0.2}, 2010_{0.6}]$ » est un intervalle de temps incertain, avec « $a^{-} = 0.2$ » et « $a^{+} = 0.6$ » sont les degrés de certitude associés aux deux bornes. Les deux degrés de certitude sont représentés dans l'ontologie en utilisant « *HasBeginnigCertainty* » et « HasEndCertainty ». La première propriété a comme co-domaine (range) « 0.2 » et la seconde a comme co-domaine « 0.6 ».

3.1.2 Points de temps certains et incertains

Nous étendons l'approche 4D-fluents pour représenter des points de temps certains et incertains en introduisant une classe nommée « *TimePoint* ».

Soit P un point de temps certain. Nous introduisons une propriété de type de données *(Datatype Property*) nommée « *HasTimePoint* » qui relie la classe « *TimePoint* » à P. Par exemple, pour l'expression « Nicolas a quitté le pays en 2000 », « 2000 » est un point de temps certain lié à la classe « *TimePoint* » en utilisant la propriété de type de données « *HasTimePoint* ».

Soit Qq un point de temps incertain, où « q » est le degré de certitude associé au point de temps Q. Nous utilisons la propriété de type de données « *HasTimePoint* » pour relier la classe « *TimePoint* » à Q. Pour représenter le degré de certitude associé à Q, nous proposons une propriété de type de données certaine nommée « *PointCertainty* » associée à la classe « *TimePoint* ». Prenons par exemple l'information suivante donnée par une patiente atteinte de la maladie d'Alzheimer : « *Janette a donné naissance à son premier enfant peut-être en 2004* », « $2004_{0.9}$ » est un point de temps incertain avec un degré de certitude associé « q = 0.9 ». « 2004 » est lié à la classe « *TimePoint* » en utilisant la propriété de type de données « *HasTimePoint* » et « 0.9 » est le degré de certitude associé représenté dans l'ontologie à l'aide de la propriété de type de données « *PointCertainty* ».

3.2 Données temporelles qualitatives

Pour représenter des relations temporelles certaines et incertaines entre des intervalles de temps et des points de temps, nous avons proposé quatre propriétés d'objet. La propriété « *RelationIntervals* » qui relie deux instances de la classe « *TimeInterval* » pour représenter les relations Intervalle-Intervalle. La propriété « *RelationIntervalPoint* » qui relie une instance de la classe « *TimeInterval* » (domaine) et une instance de la classe « *TimePoint* » (range) pour représenter les relations Intervalle-Point. La propriété « *RelationPointInterval* » qui relie une instance de la classe « *TimePoint* » (domaine) et une instance de la classe « *TimeInterval* » (range) pour représenter les relations point-intervalle. Enfin « *RelationPoints* » qui relie deux instances de la classe « *TimePoint* » pour représenter les relations Point-Point.

Pour les relations temporelles qualitatives incertaines, quatre propriétés d'objet ont également été proposées pour représenter le degré de certitude associé à une relation donnée. « *RelationIntervalsCertainty* » qui relie deux instances de la classe « *TimeInterval* » pour représenter un degré de certitude associé à une relation intervalle-intervalle. « *Relation-IntervalPointCertainty* » qui relie une instance de la classe « *TimeInterval* » et une instance de la classe « *TimePoint* » pour représenter un degré de certitude associé à une relation Intervalle-Point. « *RelationPointIntervalCertainty* » qui relie une instance de la classe « *TimePoint* » et une instance de la classe « *TimeInterval* » pour représenter un degré de certitude associé à une relation point-intervalle. Et enfin « *Relation-PointsCertainty* » qui relie deux instances de la classe « *TimePoint* » pour représenter un degré de certitude associé à une relation point-point.

4 Raisonnement sur des données temporelles certaines et incertaines en OWL 2

Notre approche étend l'algèbre des intervalles d'Allen pour : *(i)* raisonner sur des données temporelles quantitatives certaines et incertaines pour inférer des relations temporelles qualitatives et *(ii)* raisonner sur les relations temporelles qualitatives pour en inférer de nouvelles relations.

par des non-experts et permettent une mise en correspondance de l'ontologie avec le modèle probabiliste grâce à son modèle graphique convivial [44]. Il fournit un modèle à base d'arcs nous permettant de représenter des ontologies [25].

Par exemple, si nous avons le degré de certitude $a^+ = 0.4$ et le degré de certitude $b^- = 0.5$ alors $c = 0.43$, comme le montre la Figure 2.

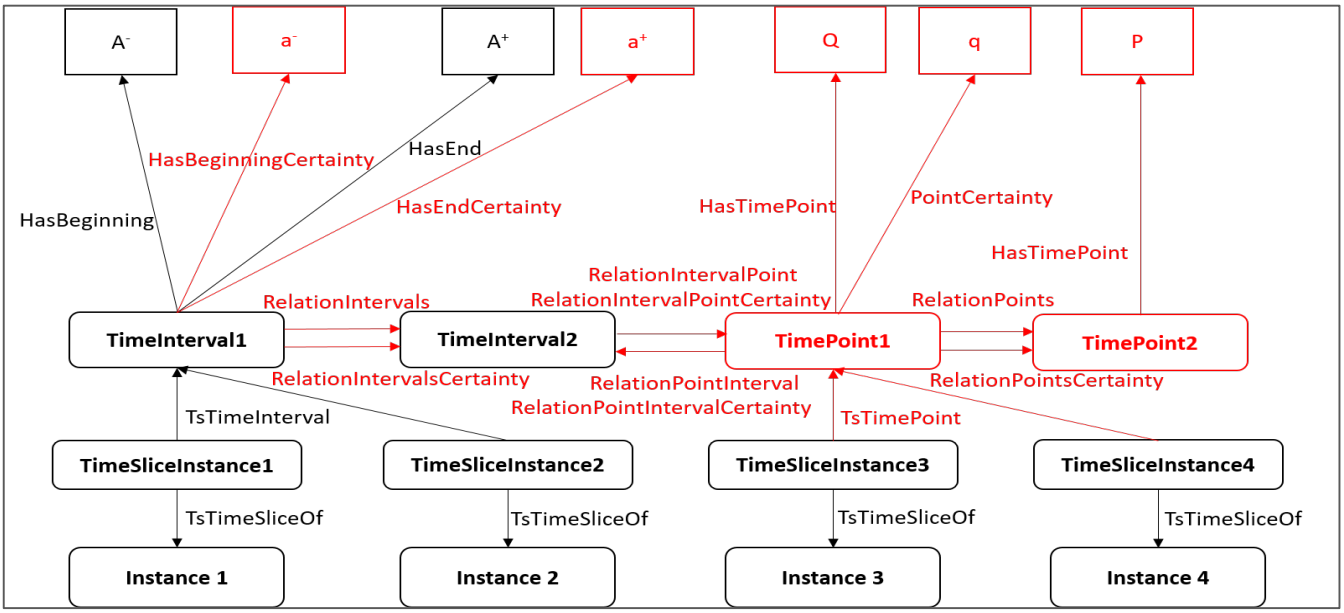


Figure 1. Notre extension de l'approche 4D-Fluents pour représenter des données temporelles certaines et incertaines

4.1 Relations entre les intervalles de temps

En considérant les intervalles de temps certains, notre approche se réduit à l'algèbre d'intervalle d'Allen. Nous redéfinissons les 13 relations d'Allen pour proposer des relations temporelles entre des intervalles de temps incertains. Soit $A = [A^-_{a-}, A^+_{a+}]$ et $B = [B^-_{b-}, B^+_{b+}]$ deux intervalles de temps incertains. Par exemple, nous redéfinissons la relation « Before(A, B) » comme : « Before_c(A, B) »; où « c » est le degré de certitude associé à la relation « *Before* » entre A et B. Cela signifie que la borne inférieure incertaine de l'intervalle A est inférieure à la borne supérieure incertaine de B.

$$\text{Before}_c(A, B) \Longrightarrow A^+_{a+} < B^-_{b-}$$

Le degré de certitude « c » est déduit des degrés de certitude a^+ et b^- en utilisant un réseau bayésien. Ce dernier représente, selon la littérature, une technique adaptée pour travailler avec l'incertitude typique des applications réelles [38], [25] et [12]. Selon [42], les réseaux bayésiens ont un pouvoir expressif et une capacité de raisonnement probabiliste rigoureuse et efficace. En outre, il représente un outil graphique puissant pour représenter, apprendre et calculer des distributions de probabilités [43]. Les réseaux bayésiens peuvent être analysés

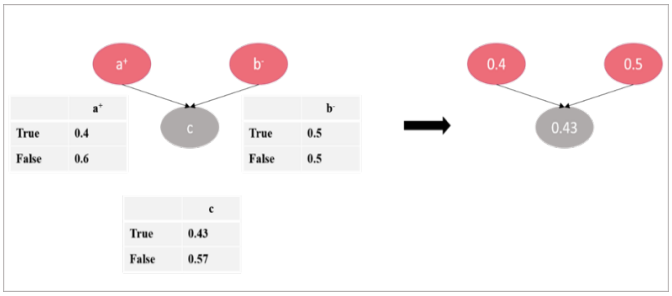


Figure 2. Réseau bayésien associé à la relation « Before_c » entre deux intervalles.

Nous classons les relations en deux catégories. La première couvre les relations qui consistent en une formule unique sans aucune conjonction (c'est-à-dire « Before_c », « after_c », « Meet_c » et « Met-by_c »). La seconde couvre le reste des relations. Celles-ci sont définies sur la base d'un ensemble de conjonctions (par exemple, « Overlaps_c »). Ainsi, comme pour la relation « Before_c », nous redéfinissons la relation « Meets_c(A, B) »; comme indiqué dans le Tableau 1.

Nous redéfinissons la relation « Overlaps(A, B) » comme suit :

$$\text{Overlaps}_c(A, B) \Longrightarrow (A^-_{a-} < B^-_{b-}) \wedge (A^+_{a+} > B^-_{b-}) \wedge (A^+_{a+} < B^+_{b+})$$

Où « c » est le degré de certitude associé à la relation « Overalps » entre A et B. Cela signifie que la borne supérieure incertaine de A est inférieure à la borne supérieure incertaine de B; la borne inférieure incertaine de A est supérieure à la borne supérieure incertaine de B; et la borne inférieure incertaine de l'intervalle A est inférieure à la borne inférieure incertaine de B.

Tableau 2. Relations temporelles entre deux intervalles de temps incertains A et B

Relation(A, B)	**Définition**	**Inverse(B, A)**
$Before_c(A, B)$	$A^+_{a+} < B^-_{b-}$	$After_c(B, A)$
$Meets_c(A, B)$	$A^+_{a+} = B^-_{b-}$	$Met\text{-}by_c(B, A)$
$Overlaps_c(A, B)$	$(A^-_{a-} < B^-_{b-}) \wedge (A^+_{a+} > B^-_{b-}) \wedge (A^+_{a+} < B^+_{b+})$	$Overlapped\text{-}by_c(B, A)$
$Starts_c(A, B)$	$(A^-_{a-} = B^-_{b-}) \wedge (A^+_{a+} < B^+_{b+})$	$Started\text{-}by_c(B, A)$
$During_c(A, B)$	$(A^-_{a-} > B^-_{b-}) \wedge (A^+_{a+} < B^+_{b+})$	$Contains_c(B, A)$
$Ends_c(A, B)$	$(A^-_{a-} > B^-_{b-}) \wedge (A^+_{a+} = B^+_{b+})$	$Ended\text{-}by_c(B, A)$
$Equals_c(A, B)$	$(A^-_{a-} = B^-_{b-}) \wedge (A^+_{a+} = B^+_{b+})$	$Equals_c(B, A)$

Le degré de certitude « c » est déduit des degrés de certitude « c1 », « c2 » et « c3 » associés respectivement aux conjonctions ($A^-_{a-} < B^-_{b-}$), ($A^+_{a+} > B^-_{b-}$) et ($A^+_{a+} < B^+_{b+}$), en utilisant quatre réseaux bayésiens. Le premier consiste à définir le degré de certitude « c1 ». Le second à définir le degré de certitude « c2 ». Le troisième à définir le degré de certitude « c3 ». Et le dernier à définir le degré de certitude « c » de la relation entre « c1 », « c2 » et « c3 », comme le montre la Figure 3.

Nous redéfinissons les autres relations d'Allen de la même manière que la relation « Overlapsc(A, B) », comme indiqué dans le Tableau 2.

Les relations proposées entre des intervalles de temps incertains préservent de nombreuses propriétés de l'algèbre d'Allen.

Réflexivité / Irréflexivité : les relations {« $Before_c$ », « $After_c$ », « $Meets_c$ », « $Met\text{-}by_c$ », « $Overlaps_c$ », « $Overlapped\text{-}by_c$ », « $Starts_c$ », « $Staretd\text{-}by_c$ », « $During_c$ », « $Contains_c$ », « $Ends_c$ » et « $Ended\text{-}by_c$ »} entre des intervalles de temps incertains sont irréflexives. Soit R une de ces relations. Il tient que : R(A, A) = 0. Par exemple :

$$Before_c(A, A) = (A^+_{a+} < A^-_{a-}) = 0 \text{ comme } A^+_{a+} > A^-_{a-}$$

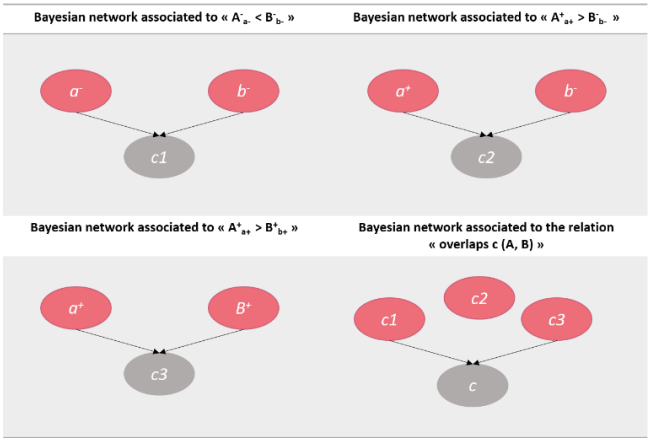


Figure 3. Réseau bayésien associé à la relation « $Overlaps_c$ » entre deux intervalles de temps incertains.

« $Equals_c$ » est réflexif. Il tient que :

$$Equal_c(A, A) = (A^-_{a-} = A^-_{a-}) \wedge (A^+_{a+} = A^+_{a+}) = 1$$

Symétrie / Asymétrie : les relations {« $Before_c$ », « $After_c$ », « $Meets_c$ », « $Met\text{-}by_c$ », « $Overlaps_c$ », « $Overlapped\text{-}by_c$ », « $Starts_c$ », « $Staretd\text{-}by_c$ », « $During_c$ », « $Contains_c$ », « $Ends_c$ » et « $Ended\text{-}by_c$ »} entre des intervalles de temps incertains sont asymétriques. Soit R une de ces relations. Il tient que : R (A, B) et R (B, A) → A = B

Par exemple, on peut déduire de « $Before_c(A, B)$ » et « $Before_c(B, A)$ » que « A = B » est vrai. En effet, « ($A^+_{a+} < B_b$) » est déduit de « $Before_c(A, B)$ », et « ($B^+_{b+} < A^-_{a-}$) » est déduit de « $Before_c(B, A)$ ». De « ($A^+_{a+} < B^-_{b-}$) » et « ($B^+_{b+} < A^-_{a-}$) », nous concluons que « A = B ».

« $Equals_c$ » est symétrique. Il tient que :

$$Equals_c(A, B) = Equals_c(B, A)$$

Transitivité : Les relations {« $Before_c$ », « $After_c$ », « $Overlaps_c$ », « $Overlapped\text{-}by_c$ », « $Starts_c$ », « $Staretd\text{-}by_c$ », « $During_c$ », « $Contains_c$ », « $Equals_c$ »} entre des intervalles de temps incertains sont transitives. Soit R une de ces relations. Soit S = [S^-_{s-}, S^-_{s+}] un intervalle de temps incertain. Il tient que:

$$R(I, J) \text{ and } R(J, S) \Rightarrow R(I, S)$$

Par exemple, nous pouvons déduire de « $Before_c(A, B)$ » et « $Before_c(B, S)$ » que « $Before_c(A, S)$ » est valide. En effet, « ($A^+_{a+} < B^-_{b-}$) » est déduit de « $Before_c(A, B)$ » et « ($B^+_{b+} < S^-_{s-}$) » est déduit de « Beforec(B, S) », par conséquent « ($A^+_{a+} < S^-_{s-}$) » est valide. Cela signifie que « $Before_c(A, S)$ » est valide.

4.2 Relations entre un intervalle de temps et un point de temps

Nous avons adapté les relations temporelles qualitatives entre les intervalles de temps pour proposer des relations entre un intervalle de temps et un point temporel, c'est-à-dire des relations intervalle-point et point-intervalle, comme indiqué dans le Tableau 3.

Tableau 3. Relations temporelles entre un intervalle de temps A et un point de temps P

Relation(P, A)	Définition	Illustration	Inverse(A, P)
Relation temporelle entre un intervalle de temps certain A = [A^-, A^+] et un point de temps certain P			
Before(P, A)	$P < A^-$		After(A, P)
After(P, A)	$A^+ < P$		Before(A, P)
Starts(P, A)	$P = A^-$		Started-by(A, P)
During(P, A)	$(A^- < P) \wedge (P < A^+)$		Contains(A, P)
Ends(P, A)	$P = A^+$		Ended-by(A, P)
Relation temporelle entre un intervalle de temps incertain A = [A^-_{a-}, A^+_{a+}] et un point de temps incertain P_p			
$Before_c(P_p, A)$	$P_p < A^-_{a-}$		$After_c(A, P_p)$
$After_c(P_p, A)$	$A^+_{a+} < P_p$		$Before_c(A, P_p)$
$Starts_c(P_p, A)$	$P_p = A^-_{a-}$		$Started\text{-}by_c(A, P_p)$
$During_c(P_p, A)$	$(A^-_{a-} < P_p) \wedge (P_p < A^+_{a+})$		$Contains_c(A, P_p)$
$Ends_c(P_p, A)$	$P_p = A^+_{a+}$		$Ended\text{-}by_c(A, P_p)$

4.3 Relations entre les points de temps

Les relations temporelles qualitatives entre les intervalles de temps ont été adaptées pour proposer des relations entre les points de temps, comme le montre le Tableau 4.

Tableau 4. Relations temporelles entre deux points de temps P et Q

Relation(P, Q)	Définition	Inverse(Q, P)
Relations temporelles entre les points de temps certains P et Q		
Before(P, Q)	$P < Q$	After(Q, P)
Equals(P, Q)	$P = Q$	Equals(Q, P)
Relations temporelles entre les points de temps incertains P_p and Q_q		
$Before_c(P, Q)$	$P_p < Q_q$	$After_c(Q, P)$
$Equals_c(P, Q)$	$P_p = Q_q$	$Equals_c(Q, P)$

Nous proposons des tables de transitivité pour dériver de nouvelles données temporelles à partir des relations temporelles qualitatives entre les intervalles de temps et les points de temps. Les quatre tables sont présentées dans notre article journal [1].

5 « UncertTimeOnto » : une ontologie des données temporelles incertaines en OWL 2

Cette section présente notre ontologie OWL 2, appelée « *UncertTimeOnto* »[2]. Elle implémente notre extension de l'approche 4D-fluents et de l'algèbre d'Allen. Nous instancions, sur la base de notre extension de l'algèbre d'Allen, les propriétés d'objet suivantes :

« *RelationIntervals* », « *RelationIntervalsCertainty* », « *RelationIntevalPoint* », « *ReltaionIntervalPointCertainty* », « *RelationPointInterval* », « *RelationPointIntervalCertainty* », « *RelationPoints* » et « *RelationPointsCertainty* ». Par exemple, « *RelationIntervals* » peut être l'une des relations d'Allen. En d'autres termes, 13 propriétés d'objet existent : « *BeforeIntervals* », « *MeetsIntervals* », « *OverlapsIntervals* », « StartsIntervals », « *DuringIntervals* », « *EndsIntervals* », « AfterIntervals », « *MetbyIntervals* », « *OverlappedbyIntervals* », « *StartedbyIntervals* », « *ContainsIntervals* », « *EndedbyIntervals* » et « *EqualsIntervals* ».

Notre ontologie comporte également un ensemble de règles SWRL pour déduire de nouvelles relations temporelles qualitatives. Pour chaque relation temporelle, nous associons une règle SWRL pour la déduire des données temporelles quantitatives fournies par l'utilisateur. Sur la base des tables de transitivité, nous associons une règle SWRL pour chaque relation de transitivité. La Figure 4 montre un exemple d'une règle SWRL.

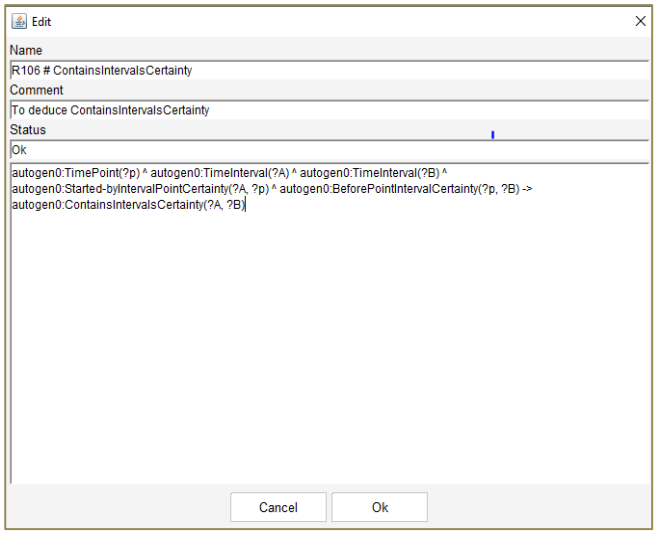


Figure 4. Exemple d'une règle SWRL

« *UncertTimeOnto* » inclut 2 classes, 6 propriétés de type de données, 64 propriétés d'objet et 180 règles SWRL. Elle peut être intégrée dans d'autres ontologies certaines ou probabilistes pour traiter des données temporelles certaines et incertaines. Nous avons utilisé, pour l'implémenter, l'éditeur d'ontologie Protégé, comme la montre la Figure 5.

[2]https://cedric.cnam.fr/isid/ontologies/UncertTimeOnto.owl#

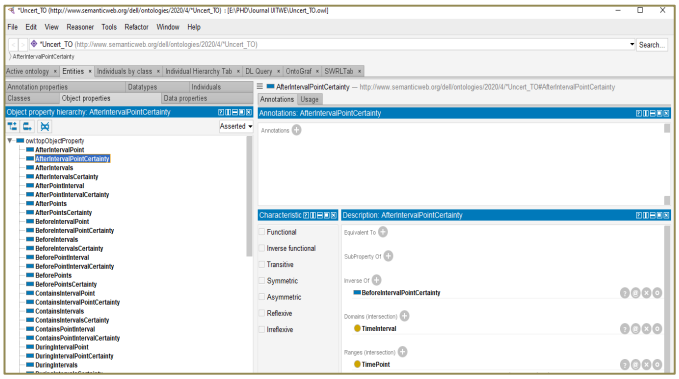


Figure 5. Notre ontologie crée par l'éditeur Protégé

6 Expérimentation

Pour valider notre approche, nous avons mené une étude de cas dont le but est de montrer l'intérêt de notre approche dans le cadre de la prothèse de mémoire CAPTAIN MEMO.

Nous avons implémenté un prototype basé sur l'ontologie « *UncertTimeOnto* ». Ensuite, nous l'avons intégré dans CAPTAIN MEMO pour gérer la dimension temporelle des données décrites par l'ontologie « *PersonLink* ».

6.1 Prototype implémenté

Pour valider notre approche, nous avons implémenté avec le langage Java, un prototype basé sur notre ontologie « *UncertTimeOnto* ». Nous avons utilisé Netica-J[3] pour implémenter les réseaux bayésiens et OWL API pour manipuler les ontologies.

Ce prototype inclut trois composants : « *UncertTimeOnto Population* », « Qualitative Temporal Data Inference » et « *Querying* ». Tout d'abord, l'utilisateur instancie « *UncertTimeOnto* » via une interface utilisateur. Cette interface permet à l'utilisateur de saisir des données temporelles certaines et/ou incertaines. Après chaque nouvelle saisie de données temporelles, le composant « Qualitative Temporal Data Inference » est automatiquement exécuté pour déduire de nouvelles données et les degrés de certitude associés. Ce composant est basé sur les règles SWRL. Le troisième composant permet aux utilisateurs d'interroger l'ontologie via des requêtes SPARQL.

6.2 Application dans CAPTAIN MEMO

Nous intégrons notre prototype implémenté dans la prothèse de mémoire CAPTAIN MEMO pour gérer les données certaines/incertaines décrites par l'ontologie « *PersonLink* ».

Par exemple, prenons l'information suivante donnée par une patiente d'Alzheimer : « *Je pense qu'Helena a vécu en Chine de 1987 à 1990. Ensuite, elle a quitté la Chine pendant quelques années puis elle est revenue peut-être en 1996. Enfin, elle s'est installée à Hong Kong en 2000* ». Soit A = [$1987_{0.6}$, $1990_{0.2}$] un intervalle de temps incertain représentant la durée de séjour d'Helena en Chine. Soit Q = $1996_{0.8}$ un point de temps incertain représentant le retour d'Helena en Chine. Soit « P = 2000 » le point de temps certain représentant l'année de séjour d'Helena à Hong Kong. La Figure 7 illustre une partie de l'ontologie « *PersonLink* » qui représente l'intervalle de temps incertain A, l'intervalle de temps incertain Q et le moment certain P.

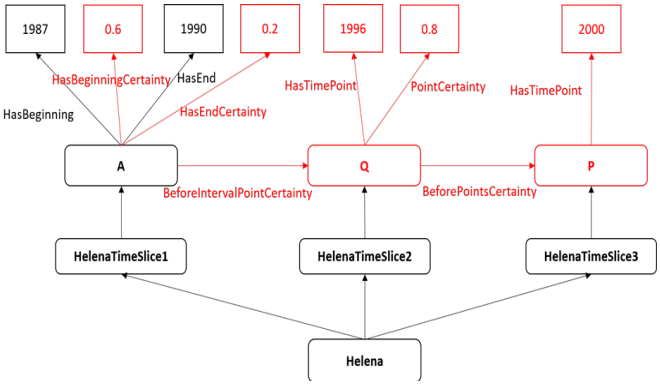


Figure 6. Un exemple d'ontologie « *PersonLink* » basée sur notre approche

7 Conclusion

Dans cet article, nous avons proposé une approche de représentation et de raisonnement sur des données temporelles certaines et incertaines en OWL 2. Cette approche comporte trois contributions. La première consiste à étendre 4D-fluents avec de nouveaux composants ontologiques « certains » pour représenter des intervalles de temps incertains et des points de temps certains/incertains ainsi que des relations temporelles qualitatives entre les intervalles de temps et les points. La seconde consiste à raisonner sur des données temporelles certaines et incertaines en étendant l'algèbre des intervalles d'Allen. Nous avons proposé 13 relations temporelles entre intervalles de temps incertains. Notre extension a l'avantage de garder les propriétés de réflexivité/irréflexivité, de symétrie/asymétrie et de transitivité. Ces relations permettent de relier un intervalle de temps et un point de temps ou deux points de temps ; où les intervalles de temps et les points peuvent être certains ou incertains. Nous avons introduit quatre tables de transitivité pour inférer les différentes relations temporelles. Enfin, la troisième contribution est la création d'une ontologie « *UncertTimeOnto* » qui implémente à la fois l'approche 4D-fluents et les extensions de l'algèbre d'Allen. Les inférences sont effectuées à l'aide de règles SWRL. Pour valider notre travail, nous avons implémenté un prototype basé sur « *UncertTimeOnto* » permettant aux utilisateurs d'explorer notre approche.

8 Références

[1] Achich, N., Ghorbel, F., Hamdi, F., Métais, E., & Gargouri, F. (2021). Certain and Uncertain Temporal Data Representation and Reasoning in OWL 2. International Journal on Semantic

[3] https://www.norsys.com/netica-j.html#download

Web and Information Systems (IJSWIS), 17(3), 51-72.

[2] Achich, N., Ghorbel, F., Hamdi, F., Metais, E., & Gargouri, F. (2019, August). Representing and Reasoning About Precise and Imprecise Time Points and Intervals in Semantic Web: Dealing with Dates and Time Clocks. In International Conference on Database and Expert Systems Applications (pp. 198-208). Springer, Cham.

[3] Allen, J. F. (1983). Maintaining Knowledge about Temporal Intervals. Commun, 26(11), 832-843.

[4] Anagnostopoulos, E. Batsakis, S., & Petrakis, E.(2013). CHRONOS: A Reasoning Engine

for Qualitative Temporal Information in OWL. Procedia Computer Science,(pp. 70-77).

[5] Artale, A., & Franconi, E. (2000). A Survey of Temporal Extensions of Description Logics. Annals of Mathematics and Artificial Intelligence, 30(1-4), 171-21.

[6] Badaloni, S., & Giacomin, M. (2006). The Algebra IAfuz: a Framework for Qualitative Fuzzy Temporal Reasoning. Artificial intelligence, 170(10), 872-908.

[7] Batsakis, S., & Petrakis, E. G. M. (2011). SOWL: A Framework for Handling Spatio-Temporal Information in OWL 2.0. In International Workshop on Rules and Rule Markup Languages for the Semantic Web (pp. 242–249).

[8] Batsakis, S., Tachmazidis, I., & Antoniou, G. (2017). Representing Time and Space for the Semantic web. International Journal on Artificial Intelligence Tools, 26(03).

[9] Baumann, R., Loebe, F., & Herre, H. (2012). Ontology of Time in GFO. In International Conference on Formal Ontologies and Information Systems (pp. 293–306).

[10] Buneman, P., & Kostylev, E. (2010). Annotation Algebras for RDFS. In Workshop on the Role of Semantic Web in Provenance Management.

[11] Ermolayev, V., Jentzsch, E., Karsayev, O., Keberle, N., Matzke, W.-E., Samoylov, V., & Sohnius, R. (2006). An Agent-Oriented Model of a Dynamic Engineering Design Process. Agent-Oriented Information Systems III (pp. 168–183). LNCS 3529.

[12] Fareh, M. (2019). Modeling incomplete knowledge of semantic web using Bayesian networks. Applied Artificial Intelligence, 33(11), 1022-1034.

[13] Freksa, C. (1992). Temporal Reasoning Based on Semi-Intervals. A.I (pp. 199–227). 54.

[14] Gammoudi A., Hadjali A., & Yaghlane B. B. (2017). Fuzz-TIME: an intelligent system for managing fuzzy temporal information. Intelligent Computing and Cybernetics, 10(2), 200-222.

[15] Ghorbel, F., Métais, E., Ellouze, N., Hamdi, F., & Gargouri, F. (2017, March). Towards accessibility guidelines of interaction and user interface design for Alzheimer's disease patients. In Tenth International Conference on Advances in Computer-Human Interactions.

[16] Ghorbel, F., Hamdi, E. & Métais, E. (2019). Ontology-Based Representation and Reasoning about Precise and Imprecise Time Intervals. In IEEE International Conference on Fuzzy Systems.

[17] Ghorbel, F., Hamdi, F., Metais, E. (2019, June). Estimating the Believability of Uncertain Data Inputs in Applications for Alzheimer's Disease´ Patients. In International Conference on Applications of Natural Language to Information Systems (pp. 208-219). Springer, Cham.

[18] Ghorbel, F., Hamdi, F., & Métais, E. (2020). Dealing with Precise and Imprecise Temporal Data in Crisp Ontology. International Journal of Information Technology and Web Engineering (IJITWE), 15(2), 30-49.

[19] Ghorbel, F., Hamdi, F., Achich, N., & Metais, E. (2020). Handling data imperfection—False data inputs in applications for Alzheimer's patients. Data & Knowledge Engineering, 130, 101864.

[20] Guesgen, H. W., Hertzberg, J., & Philpott, A. (1994). Towards Implementing Fuzzy Allen Relations. In ECAI-94 Workshop on Spatial and Temporal Reasoning (pp. 49-55).

[21] Gutierrez, C., Hurtado, C., & Vaisman, A. (2005). Temporal RDF. In European Semantic Web Conference (pp. 93-107).

[22] Harbelot, B. A. (2013). Continuum: A Spatiotemporal Data Model to Represent and Qualify Filiation Relationships. ACM SIGSPATIAL International Workshop, (pp. 76-85).

[23] Herradi, N., Hamdi, F., Métais, E., Ghorbel, F., & Soukane, A. (2015). PersonLink: an ontology representing family relationships for the CAPTAIN MEMO memory prosthesis. In International Conference on Conceptual Modeling (pp. 3-13).

[24] Herradi, N., Hamdi, F., & Métais, E. (2017). A Semantic Representation of Time Intervals in OWL 2. In KEOD (pp. 269-275).

[25] Hlel, E., Jamoussi, S., & Hamadou, A. B. (2018). A new method for building probabilistic ontology (prob-ont). In Information Retrieval and Management: Concepts, Methodologies, Tools, and Applications (pp. 1409-1434). IGI Global.

[26] Hurtado, C. &. (2006). Reasoning with Temporal Constraints in RDF. Principles and Practice of Semantic Web Reasoning, (pp. 164–178).

[27] Kim, S. K., Song, M. Y., Kim, C., Yea, S. J., Jang, H. C., & Lee, K. C. (2008, December). Temporal ontology language for representing and reasoning interval-based temporal knowledge. In Asian Semantic Web Conference (pp. 31-45). Springer, Berlin, Heidelberg.

[28] Klein, M. C. A., & Fensel, D. (2001). Ontology Versioning on the Semantic Web. In Semantic Web Working Symposium (pp. 75-91), Stanford University, California, USA.

[29] Koubarakis, M., & Kyzirakos, K. (2010). Modeling and Querying Metadata in the Semantic Sensor Web: The model stRDF and the query language stSPARQL. In The semantic web: research and applications (pp. 425-439).

[30] Lutz, C. (2003). Description Logics with Concrete Domains - A Survey. In Advances in Modal Logics (pp. 265-296).

[31] Métais, E., Ghorbel, F., Herradi, N., Hamdi, F., Lammari, N., Nakache, D., Ellouze, N., Gargouri, F. and Soukane, A., (2012). Memory prosthesis. Non-pharmacological Therapies in Dementia, 3(2), 177.

[32] Nagypál, G., & Motik, B. (2003). A Fuzzy Model for Representing Uncertain, Subjective, and Vague Temporal Knowledge in Ontologies. In OTM Confederated International Conferences "On the Move to Meaningful Internet Systems (pp. 906–923).

[33] Noy, N., Rector, A., Hayes, P., & Welty, C. (2006). Defining N-Ary Relations on the Semantic-Web. W3C Working Group Note, 12(4).

[34] O'Connor, M. J. (2011). A Method for Representing and Querying Temporal Information in OWL. Biomedical Engineering Systems and Technologies, (pp. 97-110).

[35] Preventis, A., Petrakis, E. G., & Batsakis, S. (2014). Chronos Ed: A Tool for Handling Temporal Ontologies in PROTÉGÉ. International Journal on Artificial Intelligence Tools, 23(04).

[36] Ryabov, V., & Trudel, A. (2004). Probabilistic Temporal Interval Networks. In Temporal Representation and Reasoning (pp. 64-67).

[37] Sadeghi, K. M. M., & Goertzel, B. (2014). Uncertain Interval Algebra via Fuzzy/Probabilistic Modeling. In IEEE International Conference on Fuzzy Systems (pp. 591-598).

[38] Lamma, E., Riguzzi, F., & Storari, S. (2006). Improving the K2 algorithm using association rule parameters. In Modern Information Processing (pp. 207-217). Elsevier Science.

[39] Tappolet, J., & Bernstein, A. (2009). Applied Temporal RDF: Efficient Temporal Querying of RDF Data with SPARQL. In European Semantic Web Conference (pp. 308–322).

[40] Welty, C. A., & Fikes, R. (2006). A Reusable Ontology for Fluents in OWL. In Formal Ontology in Information Systems (pp. 226-236).

[41] Zekri, A., Brahmia, Z., Grandi, F., & Bouaziz, R. (2016). τOWL: A Systematic Approach to Temporal Versioning of Semantic Web Ontologies. Journal on Data Semantics, 5(3), 141-163.

[42] Ding, Z., & Peng, Y. (2004, January). A probabilistic extension to ontology language OWL. In 37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the (pp. 10-pp). IEEE.

[43] Gu, T., Pung, H. K., Zhang, D. Q., Pung, H. K., & Zhang, D. Q. (2004). A bayesian approach for dealing with uncertain contexts (pp. 205-210).

[44] Njah, H., Jamoussi, S., Mahdi, W., & Elati, M. (2016, October). A Bayesian approach to construct Context-Specific Gene Ontology: Application to protein function prediction. In 2016 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) (pp. 1-6). IEEE.

[45] Madkour, M., Benhaddou, D., & Tao, C. (2016). Temporal data representation, normalization, extraction, and reasoning: A review from clinical domain. Computer methods and programs in biomedicine, 128, 52-68.

[46] Li, F., Du, J., He, Y., Song, H. Y., Madkour, M., Rao, G., ... & Tao, C. (2020). Time event ontology (TEO): to support semantic representation and reasoning of complex temporal relations of clinical events. Journal of the American Medical Informatics Association, 27(7), 1046-1056.

[47] Nys, G. A., Van Ruymbeke, M., & Billen, R. (2018, October). Spatio-temporal reasoning in CIDOC CRM: an hybrid ontology with GeoSPARQL and OWL-Time. In CEUR Workshop Proceedings (Vol. 2230). RWTH Aachen University.

[48] Pons, J., Billiet, C., Pons, O., & De Tré, G. (2014). Aspects of dealing with imperfect data in temporal databases. In Flexible Approaches in Data, Information and Knowledge Management (pp. 189-220). Springer, Cham.

[49] Messaoudi, W., Farah, M., & Farah, I. R. (2019). Fuzzy Spatio-Spectro-Temporal Ontology for Remote Sensing Image Annotation and Interpretation: Application to Natural Risks Assessment. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 27(05), 815-840.

Représentation des connaissances médicales temporelles au moyen d'ontologies

J. Hilbey[1,2], X. Aimé[3,2], J. Charlet[4,2]

[1] Sorbonne Université, Paris, France
[2] Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé, LIMICS, Paris, France
[3] Cogsonomy, Nantes, France
[4] Assistance Publique-Hôpitaux de Paris, Paris, France

jacques.hilbey@sorbonne-universite.fr

Résumé

La représentation des informations temporelles, pour souhaitable qu'elle soit dans les ontologies biomédicales, n'obéit pas à une approche consensuelle. Nous proposons une ontologie fondationnelle qui combine les approches dites tridimensionnelle et quadridimensionnelle afin de pouvoir suivre l'évolution d'un individu et retracer son histoire médicale. Cela nécessite, d'une part, d'associer à toute représentation d'un individu la représentation de son parcours de vie et, d'autre part, de distinguer les propriétés qui caractérisent cet individu de celles qui caractérisent son parcours de vie.

Mots-clés

Représentation des connaissances, ontologies biomédicales, représentation du temps, aspects historiques.

Abstract

The representation of temporal information, however desirable it may be in biomedical ontologies, does not follow a consensual approach. We propose a foundational ontology that combines the so-called three-dimensional and four-dimensional approaches in order to be able to track changes in an individual and to trace his or her medical history. This requires, on the one hand, associating with any representation of an individual the representation of his or her life course and, on the other hand, distinguishing the properties that characterize this individual from those that characterize his or her life course.

Keywords

Knowledge representation, biomedical ontologies, temporal representation, historical aspects.

1 Introduction

Tout être vivant subit toutes sortes de changements au cours de sa vie, de sa conception à sa mort, et passe par des épisodes de santé et de maladie. Les connaissances médicales reflètent cette dynamique temporelle : ainsi, les manuels médicaux décrivent les entités pathologiques comme des évolutions ou des ensembles d'évolutions alternatives possibles de maladies, en les accompagnant d'informations sur les signes et les symptômes, les fréquences, les stades, les degrés de sévérité, les variantes, les processus sous-jacents et les altérations physiopathologiques, l'anatomopathologie [13]. Ces considérations, qui sont banales, entraînent des difficultés importantes lorsque l'on souhaite représenter des informations médicales temporelles à l'aide d'ontologies fondées sur des logiques de description. L'une des difficultés rencontrées tient au fait que ces logiques de description n'admettent que des relations unaires ou binaires, ce qui restreint les possibilités d'indexation temporelle [4].

Dans le domaine de la détection et de l'intervention précoces dans la psychose, qui est notre centre d'intérêt particulier [1], les aspects temporels sont omniprésents, que l'on considère la période de maturation du cerveau à l'adolescence comme période critique pour l'apparition de troubles psychiatriques, l'intérêt pour les phases précoces de la psychose, la psychose débutante comme processus dynamique, l'objectif de réduction de la durée de la psychose non traitée, ou la schizophrénie comme le dernier d'une série de stades [16].

Nous abordons ici ces difficultés de représentation des aspects temporels des connaissances médicales en faisant l'hypothèse que l'engagement ontologique pris dès l'ontologie fondationnelle, dont les ontologies de domaines vont spécialiser les concepts, fait partie de la solution.

Nous présentons dans la section 2 une ontologie fondationnelle qui se veut proche d'ontologies existantes (BFO, DOLCE) mais réévalue la place accordée aux projets, l'articulation entre entités tridimensionnelles et quadridimensionnelles, et nous proposons dans la section 3 une modélisation conceptuelle des phases, des rôles, de l'histoire d'un individu.

1. `https://psy-care.fr` - L'étude clinique à laquelle donnera lieu le projet PsyCARE permettra de recueillir des informations à partir desquelles sera établi le graphe des connaissances des patients de l'étude.

2 Méthodes

2.1 La Petite Ontologie Fondationnelle

La Petite Ontologie Fondationnelle (ontoPOF) est une ontologie fondationnelle de 35 classes dont la construction repose sur deux principes :

— Un principe d'individuation au regard de l'espace et au temps, largement inspiré de Zemach [20], qui permet de distinguer cinq types d'individus : objets, projets, événements, propriétés, informations ;

— Un principe méréotopologique de division, d'assemblage, de situation des entités conduisant à distinguer : le tout, la partie, l'interface, la composition, la position.

La version d'ontoPOF que nous présentons ici [2] est complétée par une partie d'ontologie noyau simplifiée pour nos besoins.

2.2 Tridimensionnalité et quadridimensionnalité

Représenter dans une ontologie à la fois cette chaise, mon neveu Louis, le séjour de Louis à Madrid [3], le match de football auquel il a assisté sur place, semble souhaitable. Toutefois, un débat encore vif au sein de l'ontologie (au sens philosophique) concerne la question de savoir si les objets qui nous entourent ont ou non des parties temporelles [12]. Les deux pôles principaux de ce débat sont l'endurantisme et le perdurantisme [4], qui s'accordent tous deux sur le fait que les événements ont des parties temporelles, mais s'opposent sur la manière d'envisager les objets.
Pour l'*endurantisme*, les objets n'ont pas de parties temporelles et sont « tout entier présents » [5] à chaque instant, par opposition aux événements qui se déploient dans le temps. Les objets sont tridimensionnels (ou 3D) et les événements sont quadridimensionnels (ou 4D). Pour le *perdurantisme*, les objets ont eux aussi des parties temporelles. Il n'y a donc que des entités quadridimensionnelles. La position endurantiste peut s'appuyer sur le langage naturel, le « bon sens ». A l'inverse, la position perdurantiste qui considère les objets sous l'angle des « vers spatiotemporels » que constituent le déploiement de leur existence, produirait des formulations aussi contre-intuitives que « une partie temporelle de Joe est entré dans une partie temporelle de la pièce » [4].

La question qui peut se poser pour les ontologies fondationnelles est de savoir dans quelle mesure la co-présence, dans une même ontologie, d'entités de ces deux types risque de poser des problèmes de cohérence dans la modélisation.

Pour McCall et Lowe [17], la controverse 3D/4D est une tempête dans un verre d'eau, les descriptions 3D et 4D du monde étant équivalentes et traduisibles sans reste. C'est peu ou prou la position de Grenon et Smith [8], comme le rappelle Jaskolla [15] qui montre pour sa part que la traduction du quadridimensionnalisme en tridimensionnalisme pose des problèmes non résolus. Pour aller dans ce sens, si l'on considère comme McCall et Lowe un endurant, par exemple un être vivant, comme un ensemble de particules 3D, et l'équivalent 4D qu'est l'événement de sa vie, la mort de cet être diminue d'une unité le nombre d'entités dans une description 4D du monde, alors que le nombre d'entités (de particules 3D) dans une description 3D n'est pas modifié. Cette raison de privilégier une description quadridimensionnelle du monde, qui concerne l'apparition ou la disparition d'un continuant temporel, ne doit pas nous faire ignorer les aspects d'invariance que présentent certains événements, qui donnent leur force à la description tridimensionnelle.

Un argument parfois invoqué pour justifier la position d'entités tridimensionnelles est le changement : pour qu'il y ait changement, il faut qu'il y ait quelque chose qui change. Un tel argument relève de la pétition de principe, puisque c'est parce que nous avons d'abord individué un continuant temporel que nous pouvons considérer des changements le touchant. Il dénote toutefois l'importance ou la « naturalité » des descriptions tridimensionnelles.

Si nous reprenons la distinction entre objets et événements que propose une modélisation recourant à des entités aussi bien tridimensionnelles que quadridimensionnelles : on a des événements auxquels participent des objets. Mais ces objets sont eux-mêmes les lieux d'occurrence d'un certain nombre d'événements auxquels participent des objets de granularité plus fine. Ce feuilletage est repéré par Galton et Mizoguchi [6], qui voient dans l'objet une interface entre des processus [5] internes et externes, abordent l'idée d'une hiérarchie descendante de granularités éventuellement infinie, et relèvent que la dépendance ontologique des événements aux objets qui y participent doit être complétée par la dépendance ontologique des objets aux événements qui leur permettent de persister. Cette dernière considération nous semble en faveur des descriptions quadridimensionnelles.

Des points qui précèdent, il ressort que le cadre quadridimensionnel est le plus à même de permettre une représentation exhaustive du monde. Si nous conservons une branche de l'ontologie consacrée aux continuants temporels, en se conformant ainsi à la position endurantiste d'importantes ontologies fondationnelles dans le domaine biomédical [8, 7], ceux-ci ne sont pas des entités concrètes. Ce sont des « identités » stables au cours du temps de compositions très organisées de processus en interaction les uns avec les autres, susceptibles de participer comme un tout à des processus de granularité plus élevée. A ces continuants temporels doivent être associés les événements auxquels ils participent et les événements de granularité inférieure qui les constituent.

2. https://framagit.org/jacqueshilbey/pof-changement

3. L'exemple de Louis contractant une infection pendant son séjour à Madrid et recevant ensuite un traitement est discuté ci-dessous.

4. Il existe des positions mixtes que nous ne détaillerons pas ici.

5. Galton et Mizoguchi distinguent les événements des processus selon la dissectivité ; DOLCE selon la cumulativité ; BFO n'envisage que des processus. Dans POF, nous ne considérons que des événements spatiotemporels, un processus défini avant et indépendamment de sa réalisation dans un événement concret étant considéré comme un projet.

2.3 Conséquences sur la représentation du changement

En raison de l'intérêt que nous accordons à ce que nous individuons comme des continuants temporels, et en premier lieu les individus humains, il est nécessaire de tirer les conséquences du cadre que nous avons posé en matière de modélisation du changement, dans l'esprit des cas proposés par Borgo, Galton et Kutz [3].

Pour préciser la notion de changement, nous pouvons partir de la théorie des mouvements d'Aristote [1]. Plutôt que de mouvements, nous parlerions de changements puisqu'Aristote distingue la génération et la destruction, l'augmentation et la diminution, l'altération, le changement selon le lieu. C'est à ce dernier point que la langue moderne réserve le mot de mouvement.

Le déplacement physique est un événement qui n'implique pas la notion d'une modification de l'entité qui connaît ce mouvement, et dans lequel cette entité est tout entière engagée. C'est le changement typiquement invoqué à l'appui des descriptions tridimensionnelles [6, 17]. En ce sens, il est légitime de faire du continuant temporel un, voire le seul, participant à ce changement.

La génération et la destruction sont ignorées par une modélisation tridimensionnelle qui s'attache à l'identité à travers le temps de l'entité pendant son existence. Considérer l'événement de cette existence permet de modéliser la génération et la destruction comme les moments initiaux et finaux de celle-ci.

Les changements quantitatifs , on pense ici par exemple aux changements de volume, de poids, de température, font l'objet de mesures. Des propriétés d'un objet physique sont mesurées, et les valeurs de ces mesures sont susceptibles d'évoluer dans le temps. Etre individué comme un objet physique est une propriété qui n'est perdue qu'avec la destruction de l'entité. De même, les propriétés descriptives de volume, taille, température attachées à un objet physique sont des propriétés qui ne sont pas susceptibles d'être perdues par une entité matérielle : elles sont « rigides » au sens de Guarino et Welty [9]. Elles peuvent donc être attribuées au continuant temporel.

Les changements qualitatifs nécessitent une prise en compte plus fine des propriétés impliquées. S'il s'agit d'une propriété rigide, on est ramené au cas précédent (la couleur d'un objet physique, par exemple). En revanche, si nous nous intéressons à des propriétés non-rigides, par exemple au fait d'être dans une *phase* (être un adolescent), d'occuper un *rôle* (être un médecin), ou d'être dans un *état* (être malade), propriétés qui ne sont pas essentielles à toutes leurs instances, elles ne touchent pas à l'identité à travers le temps de continuants temporels mais caractérisent bien plutôt certains moments de l'événement de leur existence.

Une phase peut être naïvement définie comme une tranche temporelle d'un parcours de vie, bien que la détermination de son début et de sa fin puisse être un casse-tête [18]. Cette détermination est liée à la manière dont on la caractérise, par exemple dans le cas des *stades* d'une maladie par des événements marquant l'entrée dans cette phase ou la sortie de cette phase. Quoiqu'il en soit, ce n'est pas le continuant temporel mais bien les événements qui lui sont associés qui sont ici en jeu.

Dans la mesure où un continuant temporel est tout entier engagé dans un rôle, on pourrait être tenté d'attribuer le rôle au continuant temporel qui le tient. Toutefois, en tant que propriété non-rigide, le rôle ne ressortit pas de ce qui assure la stabilité à travers le temps de ce continuant temporel. Ce dernier point nous engage à caractériser par le rôle les événements de l'existence du continuant temporel où ce rôle est présent. Il ne peut s'agir que d'événements auxquels le continuant temporel participe. Nous proposons donc de faire porter le rôle à une spécification de la relation de participation du continuant temporel à un événement.

Enfin, un état tel qu'« être malade » appelle à nouveau une autre modélisation. Etre malade, c'est subir une évolution, la maladie, qui modifie tout ou partie des événements que nous abritons ; ce qui n'exclut pas qu'il puisse y avoir des conséquences sur les événements auxquels participe la personne malade. La maladie en tant qu'évolution est donc un événement situé spatiotemporellement, et la situation spatiotemporelle de l'événement a pour extension spatiale tout ou partie de ce qui définit en propre un continuant temporel, c'est-à-dire une certaine portion d'espace. Un état associé à un continuant temporel s'inscrit donc dans l'ensemble des événements qui constituent ce continuant temporel.

Pour représenter le « changement », la modélisation que nous proposons associe donc à un continuant temporel les propriétés rigides qui le caractérisent, ainsi que deux ensembles d'événements de granularité différente auxquels il est associé : des événements « externes » auxquels le continuant temporel participe, en spécifiant éventuellement cette relation de participation par un rôle, et des événements « internes » dont la localisation spatiotemporelle a pour extension spatiale le continuant temporel.

3 Résultats

Afin de mettre en œuvre les principes énoncés ci-dessus à propos des propriétés rigides, des rôles, des stades et des phases, nous partons d'un exemple simple : « Louis a contracté une infection lors d'un séjour à Madrid ». Cette formulation est centrée sur la principale entité tridimensionnelle impliquée, à savoir Louis. Si nous l'envisageons du point de vue des entités quadridimensionnelles impliquées, nous voyons que cette phrase articule deux événements de granularités différentes : d'une part le séjour de Louis à Madrid, événement auquel Louis participe, et d'autre part un épisode infectieux commençant lors de ce séjour à Madrid et dont Louis est le lieu d'occurrence (figure 1).

A partir de cet exemple, nous exposons trois types de connaissances d'intérêt médical comportant une dimension temporelle : l'évolution du statut de santé de Louis, l'enrichissement de son histoire médicale à partir des épisodes de maladie et de traitement dans lesquels il est impliqué, la phase de vie au cours de laquelle ces épisodes ont lieu. La représentation du temps *per se* n'est pas approfondie ici :

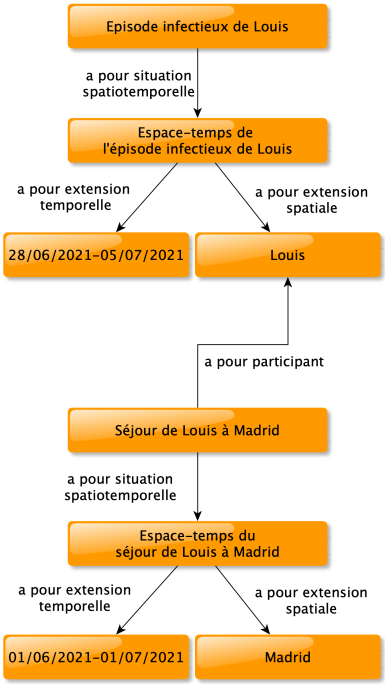


FIGURE 1 – **Un objet peut participer à un événement ou être son extension spatiale.**

nous utilisons simplement deux classes (*instant* et *intervalle temporel*) qui sont les entités temporelles de l'ontologie temporelle du W3C dans OWL [6]. Les requêtes SPARQL effectuées sur l'ontologie sont disponibles au même endroit que l'ontologie elle-même. Leur nom est indiqué dans les notes de bas de page.

Le statut de santé [7] d'un individu est une propriété dont la valeur est susceptible de changer à travers le temps ; toutefois, il est toujours possible d'attribuer cette propriété à un individu. En termes de modélisation, cela signifie qu'un individu est qualifié par la propriété « statut de santé ». Tout au long de son existence, Louis, sera qualifié par son statut propre de santé. En revanche, la valeur spécifique de ce statut de santé est susceptible de changer. Afin d'exprimer la valeur associée à un statut de santé, nous reprenons le deuxième modèle de la note du W3C *Representing Specified Values in OWL : "value partitions" and "value sets"* [8] en créant une région de valeurs spécifique aux valeurs de statut de santé qui a comme sous-classes des valeurs qui constituent une partition de cette qualité. Un statut de santé de Louis est donc une instance de deux classes : d'une part la classe générique « statut de santé » et d'autre part la sous-classe appropriée de la classe des valeurs de statut de santé. Afin de prendre en compte l'évolution temporelle du statut de santé, un statut de santé est une instance indexée temporellement à partir des *data properties* « a début » et « a fin ». Cette indexation temporelle permet de formuler une requête sur les statuts de santé qualifiant un certain individu pour les présenter par ordre chronologique et éventuellement dans un certain intervalle temporel.

L'histoire médicale [9] est une notion dont l'extension peut varier mais qui comprend au moins l'ensemble des maladies qui ont affecté un individu et des traitements qu'il a reçus. Le choix de modélisation que nous avons opéré nous amène à distinguer là aussi les deux niveaux de granularité (i) des événements auxquels participent un individu, qui définissent ce qu'on pourrait appler son « histoire externe » [10], et (ii) des événements dont un individu est le lieu, qu'on pourrait appeler son « histoire interne » [11]. Dans le domaine de la psychiatrie, plus que pour toute autre spécialité médicale, la détérioration des interactions sociales peut constituer une part prépondérante de la pathologie. Les cours évolutifs des troubles psychiatriques n'en sont pas pour autant des événements auxquels le patient participe, mais bien des événements qui ont pour siège le patient lui-même.

Reste à savoir comment sont caractérisés les événements dont un individu est le lieu et qui sont d'intérêt médical. Une partie de cette caractérisation peut provenir de la place de ces événements dans la taxonomie de l'ontologie à travers une classe appropriée (maladie, trouble) mais on peut imaginer, notamment si l'on étend l'extension du terme « histoire médicale », qu'il puisse être nécessaire de créer une propriété spécifique pour qualifier ces événements. Pour les événements auxquels un individu participe, qui sont ici les événements relatifs à un traitement, nous considérons que le rôle de patient tenu par l'individu dans ces événements doit être porté par la relation qui le lie à ces événements, comme nous l'avons vu précédemment, et que c'est par une spécification de la relation « participe à » en une relation « participe en tant que patient à » que l'on peut exprimer le rôle de patient joué par l'individu dans des événements de traitement. A partir de ces considérations, l'histoire médicale d'un individu peut être reconstituée par une requête sur l'union des événements ayant cet individu pour patient y participant et des événements ayant une situation spatiotemporelle dont l'extension est cet individu et qui sont caractérisées comme d'intérêt médical (ici parce que classées comme « maladie »).

Une phase de vie [12] est une tranche temporelle de l'existence d'un individu - en tant que membre d'une espèce - qui possède des caractéristiques développementales distinctives. Ce qui signifie que l'individu qui est actuellement dans cette phase de vie possède des caractéristiques qu'il va perdre en la quittant, tout en restant le même individu. Un exemple couramment évoqué est celui du lépidoptère qui connaît une phase « chenille » avant d'entrer dans une phase « papillon » [10, 14]. Une modélisation d'une phase de vie qui se centre sur le continuant temporel apparaît comme une contradiction dans les termes, puisqu'un continuant temporel n'a pas de parties temporelles.

La non-rigidité de la propriété, par exemple « être adolescent », nous amène à ne pas chercher à l'attribuer au conti-

6. https://www.w3.org/TR/owl-time/
7. HealthStatus.rq
8. https://www.w3.org/TR/swbp-specified-values/
9. MedicalHistory.rq
10. InternalHistory.rq
11. ExternalHistory.rq
12. PhaseOfLife.rq

nuant temporel, et donc à ne pas dire « Louis est un adolescent » mais bien plutôt à considérer les deux ensembles d'entités quadridimensionnelles associés à Louis, son histoire interne et son histoire externe, et à considérer non pas que Louis, continuant temporel, est adolescent, mais que la phase « adolescence de Louis » est une tranche temporelle des histoires de Louis. Reste, comme nous l'avons évoqué, la question de savoir quand faire commencer et quand faire finir cette phase de vie, notamment dans le cas d'une phase de vie d'un être humain comme l'adolescence, qui peut mêler dans sa définition des critères biologiques et des critères psychosociaux. Ici, nous avons choisi de nous en tenir à la définition de l'adolescence de l'Organisation Mondiale de la Santé, la période de 10 à 19 ans, et nous avons simplement défini un intervalle d'adolescence comme un intervalle temporel, et une instance d'intervalle d'adolescence comme temporellement indexée par des *data properties* « a début » et « a fin ». L'intersection de cet intervalle temporel avec les événements de l'histoire interne ou externe (c'est-à-dire les événements qui ont un début avant la fin de l'intervalle d'adolescence et une fin après le début de l'intervalle d'adolescence - en supposant donc que le début des ces événements est antérieur à leur fin), permet de reconstituer par requête les événements relatifs à l'adolescence de Louis.

4 Discussion

Dans un article qui pointe les similitudes et les différences entre BFO et DOLCE, Guarino [11] conclut sur un possible noyau commun entre les deux ontologies. Nous avons montré en quoi l'approche endurantiste devait selon nous être réaménagée. Mais dans la mesure où nous ne voulons pas proposer un modélisation conceptuelle en rupture avec le sens commun ni avec ces ontologies fondationnelles importantes dans le domaine biomédical, nous avons conservé une branche de l'ontologie consacrée aux continuants temporels. Toutefois, nous voudrions ici préciser certains points.

A la fois pour des raisons d'ergonomie cognitive (afin d'accéder plus rapidement dans l'ontologie à des classes « parlantes » pour les usagers de l'ontologie) mais aussi pour des raisons théoriquement fondées, les propriétés, les informations et les projets apparaissent au même niveau que les objets et les événements. Si dans DOLCE, les qualités [13] sont déjà au même niveau que les endurants et les perdurants, dans BFO, les propriétés aussi bien que les informations sont des continuants dépendant d'une substance, spécifiquement ou génériquement. A partir du principe d'individuation au regard de l'espace et du temps, les propriétés sont dans POF des purs continuants, sans partie temporelle ni spatiale (une chose peut être rouge pendant un temps ou sur une partie de sa surface, mais rien n'a la propriété « une partie de rouge »); les informations sont des pluricontinuants, un individu informationnel ayant certes besoin d'un support matériel pour exister, peu importe lequel (c'est ce qu'exprime la notion de dépendance ontologique *générique* dans BFO), mais pouvant aussi figurer sur différents supports à différents moments tout en restant le même individu [14]. Quant aux projets, en tant que planifications intentionnelles, ils sont dans POF des continuants spatiaux, individués selon le temps comme les objets le sont selon l'espace. Un projet peut se réaliser dans un ou plusieurs événements concrets. La comparaison que cette distinction permet, par exemple, entre une prescription et une administration de médicament, fera l'objet de travaux ultérieurs. Il résulte de ce qui précède que seuls les événements sont des entités concrètes dans POF, les autres entités présentant différents types d'abstraction, abstraction par sélection pour les propriétés et pour les objets (en tant qu'ils sont définis à partir de propriétés rigides caractérisant un agrégat de processus), abstraction par généralisation pour les informations, abstraction logique [15] pour les projets.

Les notions d'*history* et de *lifecourse* dans BFO et OGMS [16] correspondent dans leurs définitions à ce que nous avons appelé ici histoire interne et histoire externe. Notre proposition représente un approfondissement incrémental dans la mesure où des relations sont prévues pour relier un continuant temporel à ses histoires interne et externe (ce qui est rendu essentiel par le changement de perspective sur les continuants temporels), dans la mesure également où l'histoire interne est facilement constituable à partir de l'extension spatiale de la situation spatiotemporelle des événements qui la constitue, dans la mesure enfin où l'histoire externe est plus facilement décomposable à partir de relations de participation spécifiées selon le rôle.

La notion de phase apparaît dans l'OBO Foundry comme *Temporally Qualified Continuant* [14]. Jansen et Grewe relèvent son statut ontologique douteux, ce à quoi nous acquiesçons dans la mesure où par définition, un continuant n'a pas de parties temporelles. C'est pourquoi nous proposons de la définir non pas à partir du continuant temporel mais à partir des histoires interne et externe qui lui sont associées.

Nous avons considéré le statut de santé comme une propriété rigide d'un continuant temporel. La capacité à relier les modifications de valeur de cette propriété à la mention d'épisodes pathologiques de l'histoire interne fera l'objet de recherches ultérieures.

5 Conclusion

Dans cet article, nous proposons une nouvelle façon de représenter les aspects temporels des connaissances médicales, qui s'appuie pour l'essentiel sur une redéfinition des entités tridimensionnelles et quadridimensionnelles des ontologies endurantistes. La priorité est donnée aux entités quadridimensionnelles, seules entités concrètes, mais une place spécifique est reconnue - en conformité avec les

13. La distinction établie par DOLCE entre qualités et propriétés reprend la distinction d'origine aristotélicienne entre tropes et universaux.

14. donc *a minima* au niveau d'abstraction de la *manifestation* dans le modèle FRBR.

15. au sens de la méthode MERISE

16. Ontology for General Medical Science, l'ontologie noyau médicale de l'OBO Foundry

« choses » du sens commun - à la stabilité à travers le temps de certaines propriétés caractérisant des agrégats très organisés de processus (c'est l'*identité à travers le temps* des continuants temporels). Nous considérons que cette position nous permet à la fois de préserver une compatibilité forte avec les ontologies endurantistes tout en ouvrant des possibilités de modélisation pour la dynamique temporelle que présentent de nombreux phénomènes, notamment dans le domaine biomédical.

Nous avons principalement visé ici l'adéquation de la représentation. Les aspects de scalabilité et d'extensibilité seront étudiés ultérieurement, lors de l'intégration des données produites par le projet PsyCARE.

Remerciements

Ce travail a bénéficié d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du Programme d'Investissements d'Avenir portant la référence PsyCARE ANR-18-RHUS- 0014.

Références

[1] Aristote. Catégories. Paris : Éditions du Seuil ; 2002.

[2] Arp R, Smith B, Spear AD. Building ontologies with Basic Formal Ontology. MIT Press ; 2015.

[3] Borgo S, Galton A, Kutz O. Foundational ontologies in action. Applied Ontology. 2022 ;17(1) :1-16.

[4] Burek P, Scherf N, Herre H. Ontology patterns for the representation of quality changes of cells in time. J Biomed Semantics. 2019 Oct 16 ;10(1) :16.

[5] Crisp T, Smith D. 'Wholly Present' Defined. Philosophy and Phenomenological Research. 2005 Sep ;71(2) :318-344

[6] Galton A, Mizoguchi R. The water falls but the waterfall does not fall : New perspectives on objects, processes and events. Applied Ontology. 2009 Jan ;4(2) :71-107.

[7] Gangemi A, Guarino N, Masolo C, Oltramari A, Schneider L. Sweetening ontologies with DOLCE. Berlin, Heidelberg : Springer ; 2002. pp. 166–181.

[8] Grenon P, Smith B. SNAP and SPAN : Towards Dynamic Spatial Ontology. Spatial Cognition and Computation. 2004 Mar ;4(1) :69-103.

[9] Guarino N, Welty CA. A Formal Ontology of Properties. In : Dieng R, Corby O, editors. EKAW '00 : Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management ; 2000 Oct 2 ; Berlin, Heidelberg : Springer-Verlag ; c2000. p. 97-112.

[10] Guarino N, Welty CA. An overview of OntoClean. In : Staab S, Studer R, editors. Handbook on Ontologies. International Handbooks on Information Systems ; 2004 ; Berlin, Heidelberg : Springer.

[11] Guarino N. BFO and DOLCE : So Far, So Close. Cosmos + Taxis. 2017 ;4(4) :10-18.

[12] Hawley K. Temporal Parts. In : Zalta E, editor. The Stanford Encyclopedia of Philosophy. Summer 2020 ; Disponible en ligne à `https://plato.stanford.edu/archives/sum2020/entries/temporal-parts/`

[13] Hucklenbroich P. « Disease entity » as the key theoretical concept of medicine. J Med Philos. 2014 Dec ;39(6) :609-33.

[14] Jansen L, Grewe N. Butterflies and Embryos : The Ontology of Temporally Qualified Continuants. In : Ontologies and Data in Life Sciences (ODLS 2014) ; 2014 ; Universität Leipzig, Leipzig. p. E1-5.

[15] Jaskolla L. On Storms in Teacups : Limitations of 3D-4D Equivalence. Kriterion. 2011 ;25 :31-39.

[16] Krebs MO. Early detection and intervention : A change in paradigm. Ann Med Psychol (Paris). 2018 Jan ;176(1) :65-69.

[17] McCall S, Lowe E. The 3D/4D Controversy : A Storm in a Teacup. Noûs. 2006 ;40(3) :570-578.

[18] Sawyer SM, Azzopardi PS, Wickremarathne D, Patton GC. The age of adolescence. Lancet Child Adolesc Health. 2018 Mar ;2(3) :223-228.

[19] Varzi A. On the Boundary Between Mereology and Topology. In : Casati R, Smith B, White G, editors. Philosophy and the Cognitive Sciences : Proceedings of the 16th International Wittgenstein Symposium ; 1994 ; Vienna.

[20] Zemach EM. Four Ontologies. Journal of Philosophy. 1970 ;67(8) :231-247.

Session 8 : Modélisation de connaissances complexes (2)

ATLANTIS : Une ontologie pour représenter les *Instructions nautiques*

H. M. Rawsthorne[1], N. Abadie[1], E. Kergosien[2], C. Duchêne[3], E. Saux[4]

[1] LASTIG, Univ Gustave Eiffel, IGN-ENSG, F-94165 Saint-Mandé, France
[2] GERiiCO, Université de Lille, F-59000 Villeneuve d'Ascq, France
[3] LASTIG, Univ Gustave Eiffel, IGN-ENSG, F-77420 Champs-sur-Marne, France
[4] IRENav, École navale, Lanvéoc-Poulmic, CC 600, F-29240 Brest Cedex 9, France

helen.rawsthorne@ign.fr

Résumé

Les Instructions nautiques sont une série d'ouvrages produits et publiés par le Service hydrographique et océanographique de la Marine (Shom) qui donnent aux navigateurs les informations nécessaires pour naviguer près des côtes et accéder aux ports. Dans cet article, nous présentons l'ontologie ATLANTIS (coAsTaL mAritime NavigaTion InstructionS) que nous avons développée pour modéliser les connaissances contenues dans ces ouvrages, ainsi qu'un retour d'expérience et des adaptations que nous avons apportées à la Simplified Agile Methodology for Ontology Development (SAMOD), la méthodologie de développement d'ontologies que nous avons employée.

Mots-clés

Environnement maritime côtier, Instructions nautiques, navigation maritime côtière, ontologie.

Abstract

The Instructions nautiques are a series of books produced and published by the French Naval Hydrographic and Oceanographic Service, the Shom. They contain information on navigating in coastal waters and accessing ports. In this article we present the ATLANTIS (coAsTaL mAritime NavigaTion InstructionS) ontology that we have developed to model the knowledge contained in these books. We also give feedback on the Simplified Agile Methodology for Ontology Development (SAMOD), which we used to create our ontology, and present the modifications that we made to it.

Keywords

Coastal maritime environment, Instructions nautiques, coastal maritime navigation, ontology.

1 Introduction

Les *Instructions nautiques* sont une série d'ouvrages produits et publiés par le Service hydrographique et océanographique de la Marine (Shom). Elles décrivent l'environnement maritime côtier depuis le point de vue d'un navire et donnent les informations nécessaires pour naviguer près de la côte et accéder aux ports. Les *Instructions nautiques* sont utilisées par les navigateurs (civils ou militaires, professionnels ou plaisanciers) en complément des cartes marines papiers ou électroniques pour planifier leur itinéraire de navigation. Elles apportent des informations qui, pour des raisons de lisibilité, ne sont pas affichées sur une carte marine, et contribuent donc à la sécurité de la navigation maritime. Aujourd'hui, les connaissances contenues dans les *Instructions nautiques* existent uniquement sous ce forme : du texte en langage naturel. Ceci impose au Shom un processus exigeant d'actualisation et de vérification manuelle du contenu des ouvrages. Pour les utilisateurs, la forme des *Instructions nautiques* implique une lecture attentive pour trouver les informations recherchées. Sous ce forme, les connaissances sont difficilement réutilisables. Les travaux de thèse de Sauvage-Vincent [20] ont porté sur la formalisation des connaissances contenues dans les *Instructions nautiques* du Shom via la création d'un langage contrôlé. Le but du langage proposé est de « servir de pivot entre la rédaction du texte par l'opérateur dédié, la production de l'ouvrage imprimé ou en ligne, et l'interaction avec des bases de connaissances et des outils d'aide à la navigation ». Or, la base de connaissances couvrant le contenu des *Instructions nautiques* sur laquelle le langage est censé s'appuyer n'existe pas à ce jour. Nous visons donc à structurer et à géoréférencer le contenu des *Instructions nautiques* et le stocker dans une base de connaissances. À terme, cette base de connaissances pourra servir à alimenter une plateforme permettant aux utilisateurs des *Instructions nautiques* d'interroger le contenu des ouvrages. Ils pourront, au lieu de lire le texte intégral, accéder aux informations souhaitées en réalisant des recherches par facettes et notamment par zone géographique. Une première étape vers la création de cette base de connaissances consiste à se doter d'une ontologie.

Dans cet article nous présentons ATLANTIS, l'ontologie que nous avons développée pour la base de connaissances, ainsi que la méthodologie que nous avons suivie. Nous avons adopté la *Simplified Agile Methodology for Ontology Development* (SAMOD), à laquelle nous avons apporté des modifications afin de l'adapter à notre projet. Les

deux contributions de cet article sont ainsi (1) ATLANTIS, l'ontologie du domaine des *Instructions nautiques* et (2) le retour d'expériences et les adaptations de la méthodologie SAMOD. Dans la section suivante, nous présentons plus en détail les *Instructions nautiques*. Ensuite, dans la section 3, nous recensons les travaux connexes à nos recherches. Un premier état de l'art porte sur les ontologies et thésaurus existants sur le domaine maritime. Le second porte sur les méthodologies de développement d'ontologies et le choix que nous avons fait. La section 4 présente SAMOD, la méthodologie de développement d'ontologie sur laquelle nous nous sommes appuyés, et la manière dont nous l'avons mise en œuvre pour créer ATLANTIS. Enfin, dans la section 5, nous présentons les résultats et l'évaluation de l'ontologie ainsi que notre retour d'expériences et nos adaptations concernant SAMOD avant de conclure dans la section 6.

2 Les Instructions nautiques

Les *Instructions nautiques* font partie d'une gamme de produits diffusés par le Shom qui servent à la planification d'itinéraires de navigation maritime. D'autres ouvrages du Shom, plus spécialisés, viennent compléter les connaissances sur l'environnement côtier et la navigation, parmi lesquels on trouve *Feux et signaux de brume*, *Radiosignaux*, *Courants de marée* ainsi que l'*Annuaire des marées*. Ils apportent des renseignements qui sont nécessaires à la préparation d'un itinéraire adapté et sûr. Le type de navire, l'expérience du navigateur, la temporalité [1], les conditions météorologiques et les conditions océanographiques sont également à prendre en considération lors de la planification.

Les *Instructions nautiques* contiennent principalement trois types de renseignements [25] : (1) elles donnent des informations complémentaires à celles qui sont affichées sur les cartes marines comme les caractéristiques physiques (couleur, forme, taille, etc.) d'un amer [2], (2) elles recensent les informations absentes des cartes marines telles que le climat typique de la zone décrite, et (3) elles donnent des instructions ou des informations à propos de la navigation telles que les routes conseillées, les conditions d'accès aux ports ou encore les réglementations en place.

Les *Instructions nautiques* sont divisées en plusieurs volumes, par zone de couverture. Une zone de couverture peut être définie soit comme une section de trait de côte entre deux positions sur la côte, soit comme l'ensemble du trait de côte d'une île ou d'un ensemble d'îles. Chaque volume commence avec un chapitre de renseignements généraux. Le plan général du reste de l'ouvrage suit linéairement le trait de côte, chaque chapitre étant dédié à une section du trait de côte. En lisant un chapitre, le lecteur a l'impression d'être emmené le long de la côte par le rédacteur ; chaque repère, danger et autre particularité de l'environnement est décrit, et chaque mouillage, accès de port et entrée de chenal est détaillé. Les consignes mentionnent également les spécificités de la météorologie, la courantologie et la réglementation locales. Des photographies montrant les amers et les ports notables sont intercalées dans le texte. Elles illustrent également le positionnement relatif des différentes entités géographiques et doivent conforter le lecteur dans la représentation qu'il se fait de son environnement. La figure 1 montre un extrait des *Instructions nautiques*. On y trouve des instructions pour accéder au chenal de l'Île de Batz. Ces instructions font référence à des entités spatiales nommées et non-nommées telles que des balises, un clocher, des feux, une île et des tourelles. Des aides virtuelles à la navigation telles que des alignements, une route et un secteur sont également citées. Dans cet extrait, il y a des instructions distinctes pour des navires de différentes tailles. Les instructions peuvent, en outre, dépendre des conditions météorologiques (p. ex. par vents d'est), océanographiques (p. ex. à marée basse) ou temporelles (p. ex. de nuit). Le texte est accompagné d'une photo montrant le chenal ainsi que certaines entités spatiales citées dans le texte.

31

5.4.2.1.C. — Canal de l'Île de Batz, au NW (2012).

37 INSTRUCTIONS. — En venant de l'Est, on prend le chenal en suivant l'alignement à 293,3° du clocher de l'Île de Batz (chapelle **Notre-Dame de Bon Secours**) [48° 44,65′ N — 4° 00,58′ W], sur la côte Sud de l'île, par la pyramide blanche de l'**Île Pigued** (48° 43,98′ N — 3° 58,22′ W). Cet alignement n'est visible par les petits navires que jusqu'à environ 0,6 M à l'Est de la tourelle « Le Menk » (à mi-marée) et, par les navires à passerelle plus haute, jusqu'au Nord de la tourelle. Cet alignement se situe dans le secteur blanc (289,5° – 293°) du feu de la tourelle « Ar Chaden ». La route à 293,3° laisse au Nord le plateau des Duons et au Sud la tourelle « Le Menk » (48° 43,29′ N — 3° 56,70′ W), cardinale Ouest lumineuse, et la **Basse de Bloscon**.

43 Les petits navires, d'où il n'est pas possible de voir le clocher de l'Île de Batz, masqué par la végétation, peuvent se présenter légèrement à gauche de l'alignement de garde à 290° du phare de l'Île de Batz par la tourelle « Ar Chaden ».

49 On pénètre dans le chenal entre la tourelle « **Ar Chaden** » et la tourelle « **Men Guen Bras** ». Les dangers aux abords du chenal sont ensuite balisés par des marques cardinales portées par des tourelles et des balises. Un feu est implanté à l'extrémité de la longue estacade enracinée à la jetée de Roscoff, un autre marque l'extrémité du débarcadère fermant Porz Kernok à l'Est.

FIGURE 1 – Un extrait des *Instructions nautiques* [26, p. 399].

En dehors de l'usage d'un format numérique, la chaîne de production des *Instructions nautiques* a peu changé depuis les premières éditions publiées au XIXe siècle par l'ancêtre du Shom, le Dépôt des cartes et plans de la Marine [1]. Les textes sont mis à jour manuellement par le personnel du Shom, lorsque l'environnement change ou que de nouvelles informations sont rendues disponibles. Le Shom diffuse hebdomadairement un groupe d'avis aux navigateurs (GAN) en ligne [27] pour communiquer les mises à jour à appliquer à ses cartes marines et ses ouvrages nautiques.

3 État de l'art

3.1 Ontologies et thésaurus maritimes

D'autres travaux ont porté sur la création d'ontologies et de thésaurus du domaine maritime. En particulier, Malyan-

1. Une temporalité peut être quantitative (à 05h00) ou qualitative (de nuit). Elle peut être définie selon des horaires, des moments de la journée (avant le coucher du soleil), des dates (entre mars et septembre), des moments de l'année (en hiver) ou des horaires des marées (à mi-marée).

2. Un amer est un « objet remarquable situé à un endroit fixe sur la terre et pouvant être utilisé pour déterminer un emplacement ou une direction. » Traduit de [5].

kar [14] a proposé une ontologie pour l'information spatiale dans le domaine de la navigation maritime. Elle est fondée sur plusieurs sources officielles de connaissances maritimes, notamment les *United States Coast Pilot*. Cette série d'ouvrages, publiée par la *National Oceanic and Atmospheric Administration* (NOAA), constitue l'équivalent des *Instructions nautiques* aux États-Unis. Ce travail pourrait constituer une base importante pour notre cas d'application. Cependant, à notre connaissance, cette ontologie n'est pas publiée sur le Web et nous n'en avons pas retrouvé de trace en dehors des publications qui la décrivent.

Nous avons identifié trois projets de construction d'ontologies pour la sécurité maritime. Vandecasteele et Napoli [32] ont travaillé sur la construction d'une ontologie spatiale, associée à un moteur d'inférences géographiques, pour identifier automatiquement des navires suspects et leurs comportements probables afin d'améliorer la surveillance maritime. Le projet européen e-Compliance [4] a produit une ontologie sur les réglementations maritimes qui s'appliquent localement aux navires et aux ports. Liang et Zhai [13] ont construit une ontologie pour représenter des données liées sur le transport maritime. À notre connaissance, aucune de ces trois ontologies n'a été publiée sur le Web.

Les autres ontologies et les thésaurus que nous avons identifiés traitent essentiellement de deux sujets : la vie marine et l'environnement. Tzitzikas *et al*. [31] ont développé une ontologie sur les espèces marines et la biodiversité, et Leadbetter *et al*. [12] ont proposé une ontologie dédiée à la biologie marine et l'évolution de l'environnement maritime. Aucune des deux ontologies n'a été publiée sur le Web. Les ontologies *Semantic Web for Earth and Environmental Terminology* (SWEET) ont été développées par le *Jet Propulsion Laboratory* au *California Institute of Technology* sous contrat avec la NASA [19]. Cet ensemble d'ontologies porte sur les sciences de la Terre. Les ontologies *Property Space Direction* [3] et *Realm Hydrosphere Body* [4] contiennent des éléments qui nous intéressent particulièrement, tels que les phénomènes météorologiques et océanographiques. L'Agence européenne pour l'environnement (AEE) a développé le *GEneral Multilingual Environmental Thesaurus* (GEMET), une source commune de terminologie utilisée dans le cadre de l'agenda environnemental [5]. Il contient de nombreux termes désignant des types d'entités spatiales, dont des entités du domaine côtier telles que *anse* et *port*. Le Thésaurus Eau, qui a été créé par les six Agences de l'Eau, l'Onema (Office national de l'eau et des milieux aquatiques), la Direction de l'Eau et de la Biodiversité du Ministère de l'Écologie et du Développement Durable ainsi que l'Office International de l'Eau, porte sur le domaine de l'eau [6]. Il contient certains termes désignant des types d'entités géographiques du domaine de l'eau telles que *voie navigable* et *zone de mouillage*. Le *NERC Vocabulary Server* (NVS) est géré par le *British Oceanographic Data Centre* et le *National Oceanography Centre* (NOC), et est financé par le *Natural Environment Research Council* (NERC) au Royaume-Uni. Il est destiné aux sciences marines et contient notamment des thésaurus portant sur l'océanographie. Les thésaurus suivants nous intéressent particulièrement : l'*Oregon Coastal Atlas Coastal Erosion Thesaurus* [7], le *MIDA Coastal Erosion Thesaurus* [8] et le *Marisaurus Thesaurus* [9]. Ils contiennent des types d'entités géographiques du domaine maritime telles que *balise* et *épave*. Le NVS héberge également deux thésaurus qui contiennent des termes désignant des types de navires tels que *navire de pêche* et *navire militaire*. Ce sont le *World Meteorological Organisation voluntary observing ship category* [10] et le *SeaVoX Platform Categories* [11].

En l'absence d'ontologies existantes qui modélisent les consignes de navigation et l'ensemble des entités et phénomènes de l'environnement maritime côtier, nous en avons développé une et l'avons liée, autant que possible, aux thésaurus existants. Ce travail est décrit dans la partie 4.

3.2 Méthodologies de développement d'ontologies

Il existe de nombreuses méthodologies pour la création d'ontologies, qu'il s'agisse d'ontologies de domaine ou de haut-niveau. Nous cherchons à créer une ontologie de domaine pour représenter les connaissances contenues dans les *Instructions nautiques* et que des utilisateurs seraient susceptibles de rechercher. Nous avons donc besoin d'une méthodologie de création d'ontologie de domaine extensible, liée à d'autres vocabulaires, qui prenne en compte les besoins des utilisateurs, et qui n'impose pas une structure de données trop complexe à peupler de façon automatique à partir du texte des *Instructions nautiques*. Ces deux derniers critères impliquent de tester le peuplement de l'ontologie et l'interrogation des données au cours de sa construction pour s'assurer de sa bonne adéquation.

Nous avons étudié et comparé dix méthodologies de développement d'ontologies. Toutes commencent avec une phase d'acquisition de connaissances, de rédaction de documents de spécification ou de questions de compétence. L'étape d'intégration de concepts provenant d'autres ontologies ou ressources existantes permet de différencier les méthodologies car elle n'est pas effectuée au même moment : elle est prévue au début du processus de développement d'ontologies dans *Ontology Development 101* [15] tandis qu'elle est réservée à la dernière étape dans SAMOD [17, 18]. NeOn [30], SAMOD et MOMo proposent des consignes pour créer des ontologies modulaires. Certaines méthodologies, notamment OTK [29, 28], NeOn et SAMOD, explicitent une méthode pour évaluer l'ontologie produite lors de l'étape finale. METHONTOLOGY [3] recommande un guide d'évaluation publié dans une autre ressource et HCOME [10] souligne le besoin d'effectuer une

3. http://sweetontology.net/propSpaceDirection/
4. http://sweetontology.net/realmHydroBody/
5. http://www.eionet.europa.eu/gemet
6. http://thesaurus.oieau.fr/thesaurus/resource/ark:/99160/7af302a6-7518-4a8a-84a6-b8df7b595e14
7. http://vocab.nerc.ac.uk/collection/A02/current/
8. http://vocab.nerc.ac.uk/collection/A04/current/
9. http://vocab.nerc.ac.uk/collection/P21/current/
10. http://vocab.nerc.ac.uk/collection/C31/current/
11. http://vocab.nerc.ac.uk/collection/L06/current/

évaluation sans proposer de stratégie. *Ontology Development 101*, DOGMA [8, 9], DOGMA-MESS [2] et MOMo [21] n'intègrent aucune évaluation de l'ontologie produite, les rendant moins adaptées à nos besoins. DOGMA-MESS, DILIGENT [33] et HCOME sont orientées vers une application inter-organisationelle ou multi-acteur car elles permettent ou exigent le développement de l'ontologie en parallèle par les différents acteurs impliqués. Cette façon de travailler n'est pas adaptée à notre contexte, c'est pourquoi nous les avons écartées. METHONTOLOGY et OTK sont très similaires. Les deux commencent par l'acquisition de connaissances et la rédaction de documents de spécification. Elles se poursuivent en modélisant le domaine d'abord d'une manière informelle puis dans un langage formel. Enfin, les deux proposent une évaluation de l'ontologie produite. Les étapes de SAMOD et de MOMo sont également similaires : elles proposent un développement modulaire de l'ontologie qui est construite petit à petit, soit en ajoutant, à chaque itération, la modélisation d'une partie supplémentaire du domaine, soit en modélisant toutes les parties du domaine d'abord, puis en les fusionnant. NeOn se distingue des autres méthodologies présentées car elle fournit de nombreuses approches pour élaborer une ontologie ou un réseau ontologique. Elle demande aux ontologues de réaliser préalablement une analyse approfondie du projet afin de pouvoir choisir la bonne combinaison des processus et activités proposés.

Finalement, la méthodologie que nous avons choisie est SAMOD, essentiellement pour trois raisons. Premièrement, SAMOD permet de créer un premier squelette qui répond à l'essentiel des besoins puis de l'améliorer et de l'étendre de façon itérative. Deuxièmement, SAMOD implique fortement les experts du domaine. Cela est très utile pour un domaine aussi spécifique que le nôtre. Troisièmement, SAMOD intègre des phases de tests. Une des phases, fondée sur la transformation de questions informelles de compétence en requêtes SPARQL, permet de tester aussi bien le modèle que la création des données destinées à le peupler.

4 Mise en œuvre de SAMOD

SAMOD exige l'implication de deux types de personnes : des experts du domaine et des ontologues. Les experts doivent pouvoir décrire en langage naturel et d'une manière détaillée le domaine en question. Les ontologues doivent être capables de construire une ontologie en utilisant un langage formel à partir des descriptions informelles fournies par les experts du domaine.

4.1 Les sources de connaissances

En premier lieu, SAMOD conseille aux ontologues de récolter les informations à propos du domaine choisi avec l'aide des experts. Nous avons utilisé deux sources principales de connaissances sur le domaine de la navigation et de l'environnement maritime côtier : les documents de référence et les interactions avec les experts.

4.1.1 Les documents de référence du domaine

Nous avons consulté des documents produits par trois établissements phares dans le monde de la navigation maritime : le Shom, l'Organisation hydrographique internationale (OHI) et l'Association Internationale de Signalisation Maritime (AISM). En plus de la série d'*Instructions nautiques*, nous avons utilisé d'autres publications du Shom telles que l'ouvrage *Signalisation maritime* [22], l'ouvrage *Symboles, abréviations et termes utilisés sur les cartes marines papier* [24] et un document décrivant la base de données *Balisage maritime* [23]. De l'OHI, nous avons consulté l'*S-32 IHO Hydrographic Dictionary* [5] ainsi que le catalogue en ligne des objets décrits dans la norme S-57 [7]. Enfin, nous avons utilisé le manuel des aides à la navigation maritime de l'AISM : le *Navguide* [6].

4.1.2 Les connaissances d'experts du domaine

Nous avons identifié deux types d'experts du domaine des *Instructions nautiques* : les rédacteurs de ces ouvrages d'un côté, et les utilisateurs des ouvrages de l'autre, aussi appelés utilisateurs finaux. Ces deux groupes d'experts interagissent d'une manière différente avec les *Instructions nautiques* et ont donc leurs propres besoins les concernant. D'une part, les utilisateurs ont des habitudes de consultation efficace des *Instructions nautiques*, en parallèle des cartes marines et d'autres sources d'informations, pour planifier un itinéraire de navigation. D'autre part, les rédacteurs maîtrisent toute la chaîne éditoriale des *Instructions nautiques*. Nous avons ainsi organisé une série d'entretiens avec les utilisateurs et deux réunions avec les rédacteurs. Ces interactions sont détaillées dans la section 5.2.1.

4.2 La production de documentation

SAMOD conseille ensuite de choisir une sous-partie du domaine à modéliser. Le sous-modèle correspondant est appelé un *modelet*. Les ontologues et les experts du domaine travaillent ensemble afin de rédiger un argumentaire à propos de la sous-partie du domaine. Un argumentaire donne une description en langage naturel du problème à traiter par le modelet ainsi qu'un ou plusieurs exemples illustratifs. L'argumentaire doit être rédigé en utilisant le vocabulaire employé dans le domaine. À partir de l'argumentaire, les ontologues et les experts rédigent une liste de questions informelles de compétence. Ces questions, rédigées en langage naturel, représentent les besoins exprimés dans l'argumentaire et auxquels l'ontologie doit permettre de répondre. Pour chaque question, il faut également préciser le type de réponse attendue et fournir quelques exemples. Le dernier document à produire est un glossaire des termes utilisés dans cette sous-partie du domaine.

Nous avons préparé quatre modelets correspondants aux sous-domaines identifiés dans les *Instructions nautiques* : (1) les navires, (2) les consignes de navigation et les règlements, (3) les entités pérennes de l'environnement maritime côtier et les relations spatiales, et (4) les temporalités et les phénomènes météorologiques et océanographiques. Par la suite, nous allons prendre comme exemple le modelet (3).

Les extraits des *Instructions nautiques* des figures 1 et 6

illustrent la manière dont les entités pérennes de l'environnement maritime côtier et leurs relations spatiales sont décrites. La figure 2 montre un extrait de l'argumentaire que nous avons rédigé pour ce modelet, à l'aide des documents de référence décrits dans la section 4.1.1 et après avoir analysé et synthétisé l'utilisation et l'importance de ces concepts et de ces propriétés dans les textes des *Instructions nautiques*. Il contient le nom du modelet, une description expliquant la thématique dans le cadre des *Instructions nautiques* ainsi que son importance, et des extraits des *Instructions nautiques* qui montrent la manière dont la thématique se manifeste dans le texte.

Nom : Entités pérennes de l'environnement maritime côtier et relations spatiales
Description : Les *Instructions nautiques* contiennent des références à des entités pérennes de l'environnement maritime côtier et à des relations spatiales entre elles. Une entité pérenne de l'environnement maritime côtier peut être une entité géographique ou une aide virtuelle à la navigation. Les entités pérennes de l'environnement maritime côtier sont citées dans les *Instructions nautiques* dans les descriptions de paysages, dans les consignes de navigation, dans les règlements et dans les descriptions de phénomènes météorologiques et océanographiques. Les entités pérennes de l'environnement maritime côtier peuvent être utilisées pendant une navigation pour se situer localement...
Exemple 1 : « Le **canal de l'Île de Batz** est un chenal profond de 0,4 à 10 m qui sépare l'Île de Batz de la côte et donne accès aux ports d'échouage de **Porz Kernok** (Île de Batz) et de Roscoff. » [26, p. 398]
Exemple 2 : « La **pointe de Bloscon** porte la **chapelle de Sainte-Barbe** (48° 43,53' N — 3° 58,28' W), de couleur blanche. Plus à l'Ouest, on voit le clocher de Roscoff (48° 43,59' N — 3° 59,16' W) et, juste au Sud, les installations portuaires de Roscoff-Bloscon. » [26, p. 399]

FIGURE 2 – Extrait de l'argumentaire du modelet.

Nos questions informelles de compétence correspondent aux types d'informations recherchées par les utilisateurs des *Instructions nautiques*, d'après les résultats des entretiens mentionnés dans la section 4.1.2. La figure 3 en montre trois exemples pour le modelet « Entités pérennes de l'environnement maritime côtier et relations spatiales ». Nous avons ensuite rédigé le glossaire (cf. figure 4) dans lequel nous expliquons les termes du domaine, comme par exemple « Alignement » et « Marque latérale », grâce aux définitions données dans les documents de référence. Nous donnons également nos propres définitions pour les concepts que nous avons créés dans le cadre du modelet, tels que « Lieu de stationnement ».
À partir de l'argumentaire, des questions informelles de compétence et du glossaire, SAMOD indique que les ontologues commencent ensuite à travailler sans les experts du domaine afin de créer le modelet dans un langage formel.

4.3 Structure et formalisation d'un modelet

Un extrait du graphe du modelet « Entités pérennes de l'environnement maritime côtier et relations spatiales » est présenté dans la figure 5. Ce modelet contient trois classes principales. La première, `gsp:Feature`, est une

Question 1 : Comment est le phare de l'Île de Batz ?
Type de réponse attendu : Une liste des caractéristiques physiques du phare de l'Île de Batz et des relations spatiales le concernant.
Exemple de réponse : Le phare de l'Île de Batz est de couleur grise, haut de 43 mètres, équipé d'une balise d'émission AIS, entouré d'un groupe de maisons.

Question 2 : Où se situe le clocher de Roscoff ?
Type de réponse attendu : Les coordonnées géographiques du clocher de Roscoff et une liste des relations spatiales le concernant.
Exemple de réponse : Les coordonnées géographiques du clocher de Roscoff sont 48° 43,59' N — 3° 59,16' W. Le clocher de Roscoff est situé à l'ouest de la chapelle de Sainte-Barbe et au nord des installations portuaires de Roscoff-Bloscon.

Question 3 : Quels amers sont visibles sur l'Île de Batz ?
Type de réponse attendu : Une liste des amers qui sont sur l'Île de Batz.
Exemple de réponse : L'Île de Batz porte le phare de l'Île de Batz, la chapelle de Notre-Dame de Bon Secours, le clocher de l'Île de Batz, un sémaphore et la tour du sémaphore.

FIGURE 3 – Trois exemples de questions informelles de compétence du modelet.

Alignement : « Ligne droite définie par deux ou plusieurs amers clairement indiqués sur une carte, le long de laquelle un navire peut faire route en toute sécurité, pour entrer dans une passe, parer un danger, etc. » [5]. Un alignement est une aide virtuelle à la navigation.
Lieu de stationnement : Un lieu de stationnement est un point ou une zone géographique où un navire peut stationner temporairement ou d'une manière permanente soit en mouillant (jeter l'ancre du navire), soit en amarrant (attacher le navire à un amarrage), soit en échouant (laisser le navire toucher le fond à marée basse). Un lieu de stationnement peut être défini par la position d'un corps-mort, un dispositif d'amarrage, une échouage, un mouillage, un port, un port d'échouage, un port de pêche, un port de plaisance, une posée, une zone d'échouage, une zone de mouillage ou une zone de posées.
Marque latérale : « Les marques latérales, dont l'emploi est associé à un « sens conventionnel de balisage », sont généralement utilisées pour des chenaux bien définis. Ces marques indiquent les côtés bâbord et tribord de la route à suivre. Lorsqu'un chenal se divise, une marque latérale peut être utilisée pour indiquer la route qu'il convient de suivre de préférence (chenal préféré). Les marques latérales diffèrent suivant qu'elles sont employées dans l'une ou l'autre des régions de balisage A et B. » [22]. Dans la région A, une marque de bâbord est de couleur rouge et une marque de tribord est de couleur verte. Dans la région B, les couleurs sont inversées.

FIGURE 4 – Extrait du glossaire du modelet.

classe issue de l'ontologie GeoSPARQL [12]. GeoSPARQL est un standard de représentation et d'interrogation d'entités géographiques. Ce standard nous permet de définir la géométrie des entités et d'accéder à leurs descriptions (en WKT ou GML) par des propriétés. Nous avons identifié deux types distincts d'entités pérennes de l'environnement maritime côtier : des entités physiques,

12. http://www.opengis.net/ont/geosparql#

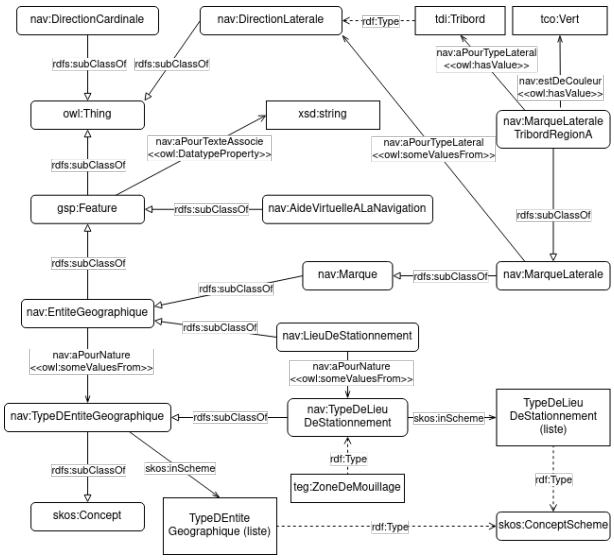


FIGURE 5 – Un extrait du graphe du modelet.

visibles, et des entités virtuelles, telles que des alignements, des dispositifs de séparation du trafic et des secteurs de feux. La classe `gsp:Feature` est donc divisée en deux sous-classes : `nav:EntiteGeographique` et `nav:AideVirtuelleALaNavigation`. Les deux autres classes principales, `nav:DirectionCardinale` et `nav:DirectionLaterale`, sont dédiées à contenir les instances permettant de décrire les caractéristiques d'entités telles que les marques latérales (cf. figure 4) et les directions de phénomènes météorologiques et océanographiques tels qu'un vent ou un courant.

Nous avons décidé de créer un thésaurus SKOS pour les entités géographiques physiques : `nav:TypeDEntiteGeographique`. Ce choix a été fait parce que le formalisme d'un thésaurus SKOS est plus simple que le formalisme descriptif rigoureux des ontologies définies avec OWL, et ne nécessite pas une description formelle de la sémantique [16]. Cela signifie qu'il sera plus facile d'enrichir un thésaurus par de nouveaux concepts SKOS que de créer automatiquement de nouvelles classes lors du peuplement de la base de connaissances. De plus, les éléments décrivant les types d'entités géographiques ne nécessitent pas l'ajout d'axiomes en plus de leurs relations hiérarchiques. Dans ce thésaurus nous avons créé trois sous-thésaurus : `nav:TypeDAmer`, `nav:TypeDeDanger` et `nav:TypeDeLieuDeStationnement`. Ils servent à rassembler les groupes d'entités géographiques physiques que nous pourrions vouloir isoler dans une requête.

Grâce aux entretiens avec des utilisateurs des *Instructions nautiques*, nous avons appris qu'ils cherchent en priorité quatre éléments principaux. Les informations sur les communications à réaliser sont recherchées pour savoir qui contacter, à quel moment et comment (numéro de téléphone ou canal VHF). Les photographies, surtout d'amers, sont fortement recherchées afin d'avoir une meilleure représentation de ce qu'il faut s'attendre à voir à l'horizon et de ce à quoi ressemble le paysage côtier en réalité. Les informations sur les ports, surtout sur les approches et les entrées, sont très recherchées ainsi que les informations administratives, celles sur le pilotage ou encore celles plus spécifiques à chaque port. Enfin, également très recherchés sont les amers, les alignements (cf. figure 4) et les feux. En cherchant des informations à propos, par exemple, d'un amer, les utilisateurs cherchent plutôt l'ensemble des caractéristiques le décrivant. Avant de consulter les *Instructions nautiques*, ils ne savent pas a priori quelles informations seront disponibles. Nous répondons à ce besoin spécifique en créant une hiérarchie dans les propriétés d'objets afin de pouvoir récupérer par inférence toutes les caractéristiques d'une instance de la classe `gsp:Feature`. Le propriété d'objets `nav:aPourCaracteristique` a donc plusieurs sous-propriétés telles que `nav:aPourForme` et `nav:estDeCouleur`.

Pendant les réunions avec les rédacteurs des *Instructions nautiques*, nous avons pris conscience qu'il est vital, pour des raisons de sécurité, de conserver les indications d'importances relatives mentionnées dans le texte. La phrase « Vue du Nord, l'Île de Batz montre la tour du sémaphore (48° 44,78' N — 4° 00,69' W) et surtout le phare (48° 44,72' N — 4° 01,61' W), tour grise haute de 43 m, entourée de maisons. » [26, p. 398] illustre ce problème. La hiérarchie indiquée par le mot « surtout » permet d'aider un navigateur dans des conditions de mauvaise visibilité en lui conseillant de privilégier, comme amer, le phare plutôt que la tour du sémaphore. Afin de résoudre ce problème, et d'autres problèmes similaires possibles liés à des nuances subtiles mais importantes dans le texte, nous avons décidé d'associer à chaque instance la phrase initiale qui la mentionne dans le texte. La propriété de données `nav:aPourTexteAssocie` a donc comme domaine `gsp:Feature` et comme range `xsd:string`.

Nous avons également créé des axiomes permettant de classer automatiquement certaines entités selon leurs propriétés et d'inférer de nouvelles connaissances. Nous allons prendre comme exemple la classe `nav:MarqueLateraleTribordRegionA` qui est destinée à stocker les instances de marques latérales tribord (cf. figure 5). Dans les *Instructions nautiques*, les marques peuvent être aussi bien désignées (1) uniquement par leur nature physique (balise, bouée, espar, tourelle) et leur type (cardinal, eaux saines, danger isolé, latéral, spécial) que (2) en utilisant le terme « marque » avec une nature physique et un type. À titre d'exemple, l'extrait « Cette route laisse dans l'Ouest la balise latérale tribord » [26, p. 320] correspond au cas (1) et l'extrait « la tourelle « Grand Pot de Beurre » (48° 37,22' N — 4° 36,47' W), marque latérale bâbord du Grand Chenal » [26, p. 413] correspond au cas (2).

Dans notre ontologie, nous avons donc déclaré la classe `nav:MarqueLateraleTribordRegionA` comme :

```
nav:MarqueLateraleTribordRegionA owl:
    equivalentClass [ owl:intersectionOf ( [
     a owl:Class ;
    owl:unionOf ( nav:Balise
```

```
                nav:Bouee
                nav:Marque
                nav:Tourelle ) ]
    [ a owl:Restriction ;
    owl:onProperty nav:aPourTypeLateral ;
    owl:hasValue tdi:Tribord ]
    [ a owl:Restriction ;
    owl:onProperty nav:estDeCouleur ;
    owl:hasValue tco:Vert ] ) ;
    a owl:Class ] .
```

De cette manière, toute entité classée comme une instance de la classe `nav:MarqueLateraleTribordRegionA` sera inférée, par exemple, comme étant de couleur verte (cf. figure 4). Inversement, toute entité classée comme `nav:Bouee`, par exemple, ayant une propriété `nav:aPourTypeLateral` pointant vers l'instance `tdi:Tribord` et une autre propriété `nav:estDeCouleur` pointant vers l'instance `tco:Vert`, sera inférée comme faisant partie de la classe `nav:MarqueLateraleTribordRegionA`.
Puis, nous avons aligné manuellement notre ontologie avec l'ensemble des ontologies et thésaurus maritimes cités dans la section 3.1 qui ont été publiés sur le Web, ainsi que des ontologies plus génériques telles que la *Spatial Relations Ontology* [13] de l'*Ordnance Survey* au Royaume Uni et l'*Extent module* de l'ontologie de haut-niveau PROTON [14] destinée à être utilisée dans la gestion des connaissances et les applications du Web sémantique.

5 Tests, résultats et discussion

La dernière étape de SAMOD consiste à tester les modelets produits, d'abord avec un test de modèle, puis avec un test de données et finalement avec des tests de requêtes. Pour faire le test de modèle, nous avons utilisé un raisonneur pour vérifier la cohérence globale de chaque modelet. Le test de données consiste à vérifier la validité du modèle après son peuplement à l'aide de triplets d'instances. Enfin, pour les tests de requêtes, il faut transformer les questions informelles de compétence en requêtes SPARQL afin de s'assurer de l'obtention des réponses attendues. Le modelet doit être ajusté jusqu'à ce que tous les tests soient concluants. Cette section présente des exemples de tests et revient sur les adaptations réalisées par rapport à SAMOD.

5.1 Les tests de modèle, de données et de requêtes

Nous avons effectué des tests de modèle sur chaque modelet et sur l'ontologie fusionnée pour tester leur cohérence en utilisant le raisonneur HermiT 1.4.3.456 intégré au logiciel Protégé.
Les amers de l'Île de Batz, ainsi que les principaux dangers qui entourent l'île, sont décrits sur la figure 6. Nous avons modélisé en triplets RDF et RDF-star une partie de cet extrait pour instancier l'ontologie (cf. figure 7, où l'entité `ent:4004` correspond au phare de l'Île de Batz). RDF-star est une extension de RDF qui permet de créer des triplets dont le sujet ou l'objet est lui-même un triplet [15]. Le fait qu'il ait été possible de modéliser des textes grâce à notre ontologie assure la validation du test de données.
Un test de requête consiste à transformer une question informelle de compétence en requête SPARQL, l'exécuter et, si le test est probant, obtenir les réponses attendues. Considérons la question 3 présentée sur la figure 3 : « Quels amers sont sur l'Île de Batz ? ». La requête SPARQL ci-dessous identifie l'entité spatiale « Île de Batz » et puis sélectionne les entités spatiales étant des amers ayant une relation spatiale `geof:sfContains` [16] avec l'île :

```
SELECT DISTINCT ?typeDAmer ?labelAmer
WHERE {
    ?entite nav:aPourNature teg:Ile.
    ?entite rdfs:label ?label.
    FILTER (regex(str(?label), "Batz"))
    ?entite geom:hasGeometry ?GeomIdB.
```

13. http://data.ordnancesurvey.co.uk/ontology/spatialrelations/
14. http://www.ontotext.com/proton/protonext

01 **5.4.2. De Roscoff à l'anse de Kernic**

01 **5.4.2.1. Île de Batz et canal de l'Île de Batz**

07 Vue du Nord, l'**Île de Batz** montre la tour du sémaphore (48° 44,78′ N — 4° 00,69′ W) et surtout le phare (48° 44,72′ N — 4° 01,61′ W), tour grise haute de 43 m, entourée de maisons. L'île est débordée de tous côtés par des dangers. Les plus au large sont :
– à l'Est, la **Basse Astan**, couverte de 0,8 m d'eau, marquée par la bouée « Astan » (48° 44,91′ N — 3° 57,66′ W), cardinale Est lumineuse ;
– au Nord, la **Grande Basse** (48° 45,92′ N — 4° 01,64′ W), couverte de 0,5 m d'eau ;
– à l'Ouest, **Men Aodi** (48° 44,64′ N — 4° 03,39′ W), roche non balisée découvrant de 0,1 m.

13

5.4.2.1.A. — Phare de l'Île de Batz, au SW (2012).

FIGURE 6 – Un extrait des *Instructions nautiques* [26, p. 398].

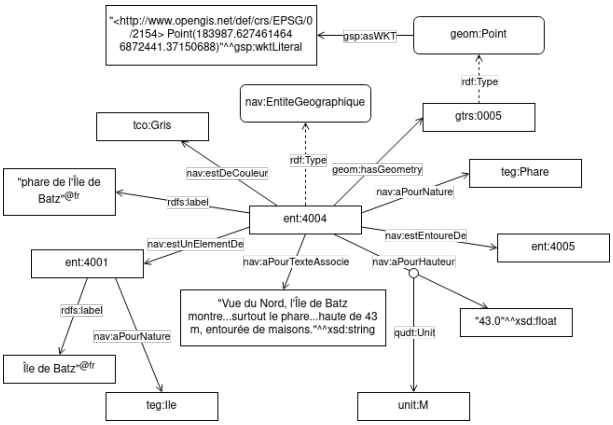


FIGURE 7 – Un graphe de quelques triplets représentant l'entité correspondant au phare de l'Île de Batz.

15. https://w3c.github.io/rdf-star/cg-spec/2021-12-17.html
16. Une fonction de GeoSPARQL : http://www.opengis.net/def/function/geosparql/sfContains

```
?GeomIdB gsp:asWKT ?WKTIdB.
?amer nav:aPourNature ?typeDAmer.
?typeDAmer a nav:TypeDAmer.
?amer geom:hasGeometry ?GeomAmer.
?GeomAmer gsp:asWKT ?WKTAmer.
OPTIONAL {?amer rdfs:label ?labelAmer}.
FILTER (geof:sfContains(?WKTIdB, ?
    WKTAmer)).}
```

Le tableau 1 présente les résultats de cette requête qui a identifié dans la base de connaissances cinq amers sur l'Île de Batz. Quatre d'entre eux sont cités dans l'extrait des *Instructions nautiques* de la figure 6; l'autre est cité sur une autre page. Grâce à l'ontologie puis à cette requête, nous avons pu rassembler ces informations qui étaient dispersées dans le texte. Pour chaque amer, la requête a récupéré son type et, si disponible, son nom. Ceci correspond à un test de requête réussi.

typeDAmer	**labelAmer**
teg:Phare	"phare de l'Île de Batz"$^{@fr}$
teg:Chapelle	"chapelle Notre-Dame de Bon Secours"$^{@fr}$
teg:Clocher	"clocher de l'Île de Batz"$^{@fr}$
teg:Semaphore	
teg:Tour	"tour du sémaphore"$^{@fr}$

TABLE 1 – Résultats de la requête SPARQL.

Le processus itératif de documentation, modélisation, évaluation et amélioration des quatre modelets et de l'ontologie fusionnée est retracé ici : https://github.com/umrlastig/atlantis-ontology. L'ontologie est composée de 110 classes, 90 propriétés d'objets, 90 propriétés de données et 2190 axiomes au total.

5.2 Retour d'expériences et adaptations de SAMOD

Nous avons essentiellement suivi la méthodologie SAMOD pour le développement de notre ontologie mais y avons apporté quelques changements afin de l'adapter à notre contexte. Ils concernent notamment (1) la relation entre les ontologues et les experts du domaine, (2) le développement en parallèle des modelets et (3) la production des données servant à peupler l'ontologie.

5.2.1 La relation entre les ontologues et les experts du domaine

SAMOD conseille aux ontologues et aux experts du domaine de recueillir des informations et de rédiger ensemble l'argumentaire, la liste de questions informelles de compétence et le glossaire. N'ayant pas la possibilité de collaborer constamment avec des experts du domaine, nous avons travaillé de manière autonome en les consultant ponctuellement pour compléter nos connaissances et valider nos travaux. Pour ce faire, nous nous sommes d'abord familiarisés avec le contenu des *Instructions nautiques* ainsi qu'avec les documents de référence détaillés dans la section 4.1.1.

Pour guider utilement la modélisation de l'ontologie, la liste des questions informelles de compétences doit refléter les besoins des utilisateurs. Nous avons donc réalisé une série d'entretiens semi-directifs auprès de personnes, ayant des niveaux d'expertise différents, qui utilisent les *Instructions nautiques* dans le cadre de leur travail ou de leurs études, dans le domaine militaire ou civil. Au total, nous avons réalisé 10 entretiens allant de 30 à 60 minutes, en petits groupes ou individuellement, avec cinq élèves en deuxième année du cycle ingénieur à l'École navale, quatre instructeurs militaires de l'École navale et trois instructeurs civils de l'École nationale supérieure maritime (ENSM). Au début de chaque entretien, nous avons présenté le cadre du projet et le but de ces entretiens, de manière à être compréhensibles pour des personnes non familières avec le Web sémantique. Nous avons précisé l'objectif de notre projet, à savoir : extraire, organiser et stocker numériquement les informations contenues dans les *Instructions nautiques*, de manière à pouvoir les utiliser différemment, en dehors de leur format habituel, en les liant à des informations provenant d'autres sources. Ensuite, nous avons indiqué que les entretiens servaient à analyser les besoins des utilisateurs des *Instructions nautiques* afin de pouvoir orienter notre travail vers des solutions adéquates. Ces entretiens ont été réalisés pendant la phase de préparation de la documentation. Ils nous ont non seulement permis de compléter nos connaissances sur le domaine des *Instructions nautiques* et leur utilisation, mais également de vérifier et de faire valider par des experts la documentation produite.

Pour recueillir l'avis des rédacteurs nous avons réalisé deux entrevues avec les membres de l'équipe de rédaction des *Instructions nautiques*. Pendant ces entrevues, nous avons présenté l'approche globale de notre projet et plus précisément l'ontologie que nous avons développée. Ensuite, nous avons eu l'opportunité de discuter et de poser des questions, pour améliorer les modelets produits.

5.2.2 Le développement en parallèle des modelets

SAMOD conseille de développer un seul modelet à la fois. Les nombres de classes, d'attributs, de relations et d'individus doivent chacun être limités à 7 ± 2. Ceci correspond à la Loi de Miller, qui indique le nombre d'éléments qu'un humain peut garder dans sa mémoire à court terme. Contrairement à ces recommandations, nous avons travaillé sur le développement de tous les modelets en parallèle. Cela nous a permis d'avoir une vue complète du domaine à tout moment et ainsi de travailler sur la modélisation de chacune de ses parties de manière non pas indépendante mais plutôt complémentaire. Cette façon de travailler nous a conduits à estimer la Loi de Miller trop contraignante, car elle aurait demandé une division trop fine du domaine des *Instructions nautiques*. Pendant la modélisation, il y a eu un va-et-vient constant inter- et intra-modelets. Il a été nécessaire d'avoir des éléments en commun entre certains modelets, comme par exemple la classe `nav:EntiteGeographique`. De cette manière, il a été plus facile de créer des liens entre des éléments de différents modelets lors de leur fusion. Après avoir défini chaque modelet, nous les avons fusionnés afin de créer l'ontologie complète. Nous les avons intégrés au fur et à mesure, en réalisant à chaque fois un test de mo-

dèle, un test de données et des tests de requêtes.

5.2.3 La production des jeux de données

Pour réaliser les tests de données, SAMOD demande aux ontologues de créer pour chaque modelet un jeu de données relatif aux exemples introduits dans la documentation. Il est conseillé de réaliser cette tâche après avoir produit le modelet et après avoir fait le test de modèle. Cependant, nous avons décidé de commencer à construire un jeu de données pour chaque modelet directement après la phase de récolte d'informations et avant de produire la documentation. Pour construire les jeux de données, nous avons structuré quelques extraits de texte des *Instructions nautiques* en construisant à la main des triplets en RDF et RDF-star. Le fait d'identifier les instances et leurs relations dès le début nous a conduits à mieux connaître le domaine et nous a aidés à mieux structurer notre modèle ontologique. Plus précisément, ceci nous a aidés à (1) identifier les sous-domaines à représenter par un modelet, (2) rédiger l'argumentaire et les questions informelles de compétence, (3) identifier les classes et propriétés nécessaires dans chaque modelet et (4) structurer le modèle global d'une manière qui permet de répondre aux usages auxquels on la destine, notamment répondre aux questions des navigateurs qui préparent leur sortie en mer. Nous avons ensuite élargi les jeux de données tout au long du développement des modelets.

6 Conclusions et perspectives

Nous avons produit l'ontologie ATLANTIS qui modélise le contenu des *Instructions nautiques* du Shom. Ce modèle permet de décrire les entités pérennes de l'environnement maritime côtier, les phénomènes météorologiques et océanographiques qui peuvent se produire dans cet environnement, les consignes de navigation côtière, les règlements en vigueur dans le domaine maritime côtier, les relations spatiales, les temporalités et enfin les navires.

Pour créer l'ontologie nous nous sommes fondés sur SAMOD, une méthodologie agile et itérative pour le développement d'ontologies de domaine. Nous lui avons apporté quelques modifications afin de l'adapter à notre contexte. Cela nous a permis de travailler directement à partir de données réelles sur un domaine très large sans devoir le découper trop finement. De plus, nous avons pu travailler de manière autonome et orienter les interactions avec les experts du domaine vers le raffinement du modèle tout en tenant compte des besoins des utilisateurs finaux de la base.

La base produite offre de nouvelles possibilités concernant à la fois l'accès aux connaissances contenues dans les *Instructions nautiques* et leur fiabilité. À terme, elle pourra alimenter une plateforme permettant aux utilisateurs des *Instructions nautiques* de chercher directement les informations précises souhaitées, par catégorie ou par zone géographique, au lieu de lire le texte intégral.

En perspectives à court terme, sur la base du peuplement de l'ontologie avec des triplets RDF et RDF-star extraits manuellement des *Instructions nautiques*, nous prévoyons d'automatiser l'extraction des connaissances géoréférencées du texte, puis de peupler la base de connaissances. Nous comptons nous appuyer sur la méthode d'extraction automatique initiée par Lamotte *et al.* [11], fondée sur une combinaison d'une approche lexicale et d'une approche à base de patrons linguistiques. Une fois le modèle ontologique instancié, il sera possible de réaliser une évaluation rigoureuse de l'ontologie produite pour ce cas d'usage.

Remerciements

Ce travail est co-financé par le Shom et l'IGN et réalisé au LASTIG, une unité de recherche de l'Université Gustave Eiffel. Nous remercions les utilisateurs des *Instructions nautiques* qui ont accepté de s'entretenir avec nous ainsi que les rédacteurs des *Instructions nautiques* au Shom.

Références

[1] G. Bessero and H. Richard, editors. *300 ans d'hydrographie française*. Locus Solus, Châteaulin, 2020.

[2] A. de Moor, P. De Leenheer, and R. Meersman. DOGMA-MESS: A Meaning Evolution Support System for Interorganizational Ontology Engineering. In *Proceedings of the 14th International Conference on Conceptual Structures*, volume 4068 of *Lecture Notes in Computer Science*, pages 189–202, Aalborg, Denmark, 2006. Springer.

[3] M. Ferndández, A. Gómez-Pérez, and N. Juristo. Methontology: From Ontological Art Towards Ontological Engineering. Technical Report SS-97-06, Laboratorio de Inteligencia Artificial, Facultad de Informáitica, Universidad Politécnica de Madrid, Spain, 1997.

[4] M. Hagaseth, L. Lohrmann, A. Ruiz, F. Oikonomou, D. Roythorne, and S. Rayot. An Ontology for Digital Maritime Regulations. *Journal of Maritime Research*, 8(2):7–18, 2016.

[5] Hydrographic Dictionary Working Group (HDWG). S-32 IHO Hydrographic Dictionary, 2019.

[6] International Association of Marine Aids to Navigation and Lighthouse Authorities (IALA). Navguide. Marine Aids to Navigation Manual, 2018.

[7] International Hydrographic Organization. S-57 IHO Object Catalogue.

[8] M. Jarrar and R. Meersman. Formal Ontology Engineering in the DOGMA Approach. In *Proceedings of the International Conference on Ontologies, Databases and Applications of Semantics*, volume 2519 of *Lecture Notes in Computer Science*, pages 1238–1254, Irvine, CA, USA, 2002. Springer.

[9] M. Jarrar and R. Meersman. Ontology Engineering - The DOGMA Approach. In *Advances in Web Semantics I*, volume 4891 of *Lecture Notes in Computer Science*. Springer, Berlin and Heidelberg, 2008.

[10] K. Kotis and G. A. Vouros. Human-centered ontology engineering: The HCOME methodology. *Knowledge and Information Systems*, 10:109–131, 2006.

[11] L. Lamotte, N. Abadie, É. Saux, and E. Kergosien. Extraction de connaissances pour la description de

l'environnement maritime côtier à partir de textes d'aide à la navigation. In *Revue des Nouvelles Technologies de l'Information*, volume Extraction et Gestion des Connaissances, RNTI-E-36, pages 341–348, Bruxelles, Belgium, 2020. Éditions RNTI.

[12] A. M. Leadbetter, T. Hamre, R. Lowry, Y. Lassoued, and D. Dunne. Ontologies and Ontology Extension for Marine Environmental Information Systems. In *Proceedings of the Workshop "Environmental Information Systems and Services - Infrastructures and Platforms"*, volume 679, pages 12–24, Bonn, Germany, 2010.

[13] Y. Liang and J. Zhai. Construction and Representation of Shipping Domain Ontology Based on Ontology Design Patterns. In *2018 15th International Conference on Service Systems and Service Management (ICSSSM)*, Hangzhou, China, 2018.

[14] R. Malyankar. Maritime Information Markup and Use in Passage Planning. In *Proceedings of the National Conference on Digital Government*, pages 25–32, Marina del Rey, California, USA, 2001.

[15] N. F. Noy and D. L. McGuiness. Ontology Development 101: A Guide to Creating Your First Ontology. Technical Report Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI- 2001-0880, Stanford University, Stanford, California, USA, 2001.

[16] J.-A. Pastor-Sánchez, F. J. Martínez Mendez, and J. V. Rodríguez-Muñoz. Advantages of thesaurus representation using the Simple Knowledge Organization System (SKOS) compared with proposed alternatives. *Information Research*, 14(4), 2009.

[17] S. Peroni. SAMOD: an agile methodology for the development of ontologies. 2016.

[18] S. Peroni. A Simplified Agile Methodology for Ontology Development. In *OWL: Experiences and Directions – Reasoner Evaluation*, Bologna, Italy, 2016.

[19] R. G. Raskin and M. J. Pan. Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). *Computers & Geosciences*, 31(9):1119–1125, November 2005.

[20] J. Sauvage-Vincent. *Un langage contrôlé pour les instructions nautiques du Service Hydographique et Océanographique de la Marine*. PhD thesis, Université Bretagne Loire, January 2017.

[21] C. Shimizu, K. Hammar, and P. Hitzler. Modular Ontology Modeling. *Semantic Web*, 2021.

[22] Shom. *Signalisation maritime*. Ouvrages généraux. 3e edition, 2016.

[23] Shom. Balisage maritime. Descriptif de contenu du produit externe, 2019.

[24] Shom. *Symboles, abréviations et termes utilisés sur les cartes marines papier*. 7e edition, 2019.

[25] Shom. Guide de rédaction des Instructions Nautiques du Shom. Procédure spécifique, Shom, October 2020.

[26] Shom. *Instructions nautiques. C22 : France (côtes Nord et Ouest). Du cap de La Hague à la pointe de Penmarc'h*. 2021. Version à jour au 20 octobre 2021.

[27] Shom. Groupe d'Avis aux Navigateurs en Ligne, 2022. https://gan.shom.fr/diffusion/home.

[28] Y. Sure. A Tool-Supported Methodology for Ontology-Based Knowledge Management. In H. Stuckenschmidt, E. Stubkjær, and C. Schlieder, editors, *The Ontology and Modelling of Real Estate Trasnsactions*, International Land Management Series, pages 115–126. Routledge, London, 2003.

[29] Y. Sure and R. Studer. On-To-Knowledge Methodology — Employed and Evaluated Version. Technical Report D16, Institute AIFB, University of Karlsruhe, Karlsruhe, Germany, 2001.

[30] M. C. Suárez-Figueroa, A. Gómez-Pérez, and M. Fernández-López. The NeOn Methodology for Ontology Engineering. In M. C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, and A. Gangemi, editors, *Ontology Engineering in a Networked World*, pages 9–34. Springer, Berlin, Heidelberg, 2012.

[31] Y. Tzitzikas, C. Allocca, C. Bekiari, Y. Marketakis, P. Fafalios, M. Doerr, N. Minadakis, T. Patkos, and L. Candela. Integrating Heterogeneous and Distributed Information about Marine Species through a Top Level Ontology. In E. Garoufallou and J. Greenberg, editors, *Metadata and Semantics Research*, Communications in Computer and Information Science, pages 289–301, Thessaloniki, Greece, 2013. Springer International Publishing.

[32] A. Vandecasteele and A. Napoli. Spatial Ontologies for Detecting Abnormal Maritime Behaviour. In *OCEANS 2012 MTS/IEEE Yeosu: The Living Ocean and Coast - Diversity of Resources and Sustainable Activities*, Yeosu, Republic of Korea, 2012. IEEE.

[33] D. Vrandečić, S. Pinto, C. Tempich, and Y. Sure. The DILIGENT knowledge processes. *Journal of Knowledge Management*, 9(5):85–96, 2005.

Liste des préfixes utilisés

ent: <http://data.shom.fr/id/entitegeographique/>
geof: <http://www.opengis.net/def/function/geosparql/>
geom: <http://data.ign.fr/def/geometrie#>
gsp: <http://www.opengis.net/ont/geosparql#>
gtrs: <http://data.shom.fr/id/codes/nav/geometries/>
nav: <http://data.shom.fr/def/navigation_cotiere#>
owl: <http://www.w3.org/2002/07/owl#>
qudt: <http://qudt.org/schema/qudt/>
rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
rdfs: <http://www.w3.org/2000/01/rdf-schema#>
skos: <http://www.w3.org/2004/02/skos/core#>
tco: <http://data.shom.fr/id/codes/nav/typedecouleur/>
tdi: <http://data.shom.fr/id/codes/nav/typededirection/>
teg: <http://data.shom.fr/id/codes/nav/typedentitegeographique/>
unit: <http://qudt.org/2.1/vocab/unit/>
xsd: <http://www.w3.org/2001/XMLSchema#>

Une ontologie pour organiser les données de processus biologiques: la contribution des modèles mathématiques

O. Inizan[1], V. Fromion[1], A. Goelzer[1], F. Saïs[2], D. Symeonidou[3]

[1] Université Paris Saclay, INRAE, MaIAGE
[2] Université Paris Saclay, LISN, CNRS UMR9015
[3] Université de Montpellier, INRAE, SupAgro, UMR MISTEA

olivier.inizan@inrae.fr

Résumé

La biologie est un domaine connu pour sa production massive de données. Ces données sont souvent qualifiées d'hétérogènes et de fragmentées, et les biologistes ne disposent pas d'une représentation formelle, qui, à l'échelle de l'organisme, permettrait de les représenter et de les organiser. Depuis une dizaine d'années les modèles mathématiques systémiques se sont révélés être des outils utiles pour comprendre le comportement de la cellule. Nous montrons dans ce travail qu'une ontologie construite sur les principes qui régissent la conception de ces modèles peut aider à organiser les données biologiques. Nous présentons ici un choix de concepts et relations compatibles avec les principes à l'oeuvre dans les modèles systémiques.

Mots-clés

Ontologies, modèles mathématiques, données biologiques.

Abstract

Biology is a research field well known for its huge quantity and diversity of data. These data are recognized as heterogeneous and fragmented. Biologists do not have a formal representation that, at the level of the entire organism, can help them to tackle such diversity and quatity. Recently, the systemic mathematical models have proven to be a powerful tool for understanding the bacterial cell behavior. We advocate that an ontology built on the principles that govern the design of such models, can help to organize the biological data. In this article we present a choice of concepts and relations compliant with principles at work in the systemic mathematical models.

Keywords

Ontology, Mathematical Models, Biological Data.

1 Introduction

En biologie, l'avancée récente des technologies de séquençage a permis une production rapide et peu onéreuse de données [15]. Aujourd'hui, les biologistes et bioinformaticiens manipulent une grande quantité et une grande diversité de données dites omiques (la génomique, la transcriptomique, la protéomique, la métabolomique et la métagénomique) [9]. Ces données sont principalement obtenues dans le contexte d'expérimentations conçues pour répondre à des questions précises. D'un point de vue plus général, les données produites sont hétérogènes et fragmentées [2]. Ainsi, malgré l'abondance de données disponibles pour un organisme particulier, la capacité de lier ces données entre elles représente un défi majeur [11]. Une telle démarche présente de nombreux intérêts comme celui d'élucider des mécanismes métaboliques en vue de traitements thérapeutiques [14]. Bien que la recherche en représentation des connaissances soit très active en biologie [13], il n'existe pas de représentation formelle destinée à organiser les données pour l'intégralité d'un organisme, aux échelles moléculaires. Une telle représentation permettrait de mieux exploiter tout le potentiel des données produites par les expériences. Depuis une dizaine d'années l'approche de modélisation "*cellule entière*" a montré que des modèles mathématiques systémiques représentent un outil important pour comprendre et décrire le comportement de la cellule bactérienne. Plus précisément, lorsque ces modèles sont calibrés à l'aide de données biologiques, il est possible d'identifier des principes organisateurs conduisant à la prédiction de comportements qui n'avaient pas été observés expérimentalement [10, 3]. Il existe donc un réel besoin de développer une nouvelle représentation formelle (i) qui a pour objectif de représenter de manière sémantique les liens entre données biologiques et (ii) qui s'inspire des principes à l'oeuvre dans la modélisation de processus biologiques.

Les travaux que nous présentons dans cet article ont été déjà publiés dans [8]. Dans ce travail nous décrivons les premières étapes du développement d'une représentation formelle destinée à l'organisation de données biologiques et conçue selon les concepts présents dans les modèles mathématiques. Ce travail est en cours de réalisation, les tâches effectuées sont principalement conceptuelles et nous évaluons l'ontologie sur un exemple simple. L'article est organisé comme suit : la section 2 présente l'état de l'art en relation avec ce travail et sa principale motivation. Les concepts et relations de l'ontologie sont présentés en section 3 et illustrés à travers l'exemple de la section 4. La section 5 présente les conclusions et perspectives.

2 État de l'art et motivation

Deux points de départ permettent de comprendre le travail présenté dans cet article : les ontologies BiPOm et BiPON [6, 5] et les contraintes relatives à la construction de modèles mathématiques.

2.1 BiPON et BiPOm

La biologie est un domaine où différentes communautés peuvent travailler sur les mêmes objets mais avec des buts différents. Il est donc crucial d'être en mesure d'éviter les ambiguïtés alors que l'on se réfère au même objet. Les bio-ontologies couramment utilisées, par exemple l'ontologie *Gene Ontology (GO)* [1] décrivent une hiérarchie de concepts utilisés comme vocabulaire contrôlé. D'autres bio-ontologies sont aussi utilisées à des fins d'échange de données : c'est le cas de BioPax ([2]) qui permet de décrire les voies métaboliques. En 2017 et 2020, deux ontologies au format OWL [1], BiPON [6] et BiPOm[5], ont proposé de nouveaux usages pour les bio-ontologies. Elles ont tout d'abord introduit l'approche systémique comme principe pour organiser la connaissance relative aux objets biologiques. Cette approche émane des sciences de l'ingénieur et consiste à découper un système en un ensemble de modules et sous-modules interconnectés [4]. Un module systémique est défini par ses entrées, ses sorties et la fonction qu'il remplit. Cette fonction, avec les entrées et les sorties sont regroupées dans un modèle mathématique. Ainsi, le comportement du module est représenté de façon formelle. Les auteurs de BiPON et BiPOm ont montré que la cellule bactérienne peut être considérée comme un système et organisée en modules et sous modules systémiques. Ces modules sont représentés par des concepts OWL nommés *processus biologiques*. D'autre part, BiPON et BiPOm sont des ontologies plus expressives que les bio-ontologies courantes, puisqu'en plus des concepts, relations et les axiomes OWL, elles contiennent également un ensemble de règles de Horn exprimées en SWRL [2]. Elles exploitent en effet les capacités de raisonnement fournies par la sémantique logique des axiomes OWL, et des règles SWRL déclarés afin d'inférer de nouvelles relations entre les individus. A titre d'exemple, un raisonnement sur l'ontologie BiPOM a montré qu'un vaste ensemble de processus biologiques pouvait être décrit par un ensemble restreint de concepts mathématiques.

2.2 Les contraintes et les modèles mathématiques

La figure 1.a illustre l'association entre un module systémique (ici le processus biologique) et son modèle mathématique. Nous présentons ici les contraintes qui permettent de construire de tels modèles. Afin de mieux comprendre ces contraintes, il faut d'abord détailler un peu plus le concept de processus biologique tel qu'il est défini dans les ontologies BiPON/BiPOm. Un processus biologique a une ou plusieurs molécules comme entrée et une ou plusieurs molécules en sortie. Nous dirons qu'un processus *consomme* ses entrées et *produit* ses sorties. De plus et comme évoqué ci dessus, un processus remplit une fonction. Le processus possède finalement un moyen de transformer les entrées en sorties et ce moyen est exprimé au travers d'un modèle mathématique. La forme générale d'un processus biologique est détaillée dans la figure 1a. La figure 1b présente une simple réaction biochimique (la conversion d'une molécule 'A' en molécule 'B') et le processus biologique correspondant.

Un point important est que quel que soit le modèle mathématique construit, trois contraintes sont toujours respectées. Nous considérons donc que (i) ces contraintes sont majeures et (ii) qu'elles pilotent la construction de modèles mathématiques. Dans la suite nous nommerons ces contraintes les *contraintes des modèles*. Ces trois contraintes sont :

1. *La causalité physique*. En physique la causalité indique que si les entrées d'un modèle produisent les sorties du modèle alors les entrées précèdent les sorties. Dans la représentation que nous construisons nous ne considérons pas le temps et nous reformulons la causalité ainsi : si les entrées sont présentes en quantité suffisante alors le processus peut consommer les entrées et produire les sorties.
2. *La conservation de la masse* est une contrainte importante pour la construction de modèles. Elle assure leur consistance.
3. *La compétition pour l'accès aux ressources*. Dans la cellule les processus biologiques sont en compétition pour l'accès aux ressources. Plus précisément, les mêmes molécules peuvent être consommées par des processus différents. Ainsi, la molécule d'ATP fournit l'énergie nécessaire à la cellule et est, par conséquent, consommée par de nombreuses réactions biochimiques.

Malgré le fait que le concept de processus biologique soit présent dans les ontologies BiPON/BiPOM et que les modèles mathématiques sont représentés dans BiPON, aucune de ces deux ontologies ne considère ces contraintes.

2.3 Motivation

Nous envisageons les contraintes des modèles comme un moyen de valider la consistance de la connaissance biologique et des données associées à un organisme. Si nous souhaitons utiliser ces contraintes dans une représentation formelle, nous devons d'abord fournir des concepts et des relations qui nous permettent de *compter* les molécules produites ou consommées par les processus. Si l'on considère l'exemple de la figure 1b : la causalité indique qu'il faut au moins une molécule A disponible pour produire une molécule B. La compétition pour l'accès aux ressources nécessite aussi de compter les molécules. Imaginons qu'un processus P' consomme également de la molécule A. S'il y a seulement *une seule* molécule A dans toute la cellule, P et P' sont en compétition. Comme évoqué dans la section 2.2 BiPON et BiPOm ont validé l'approche systémique pour

1. https ://www.w3.org/OWL/
2. https ://www.w3.org/Submission/SWRL/

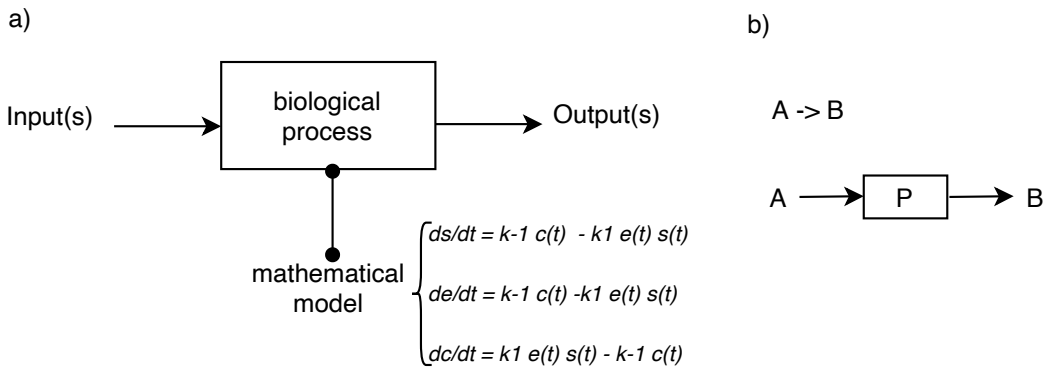


FIGURE 1 – a) Un processus biologique et le modèle mathématique associé. b) Une réaction biochimique et le processus P qui la représente.

représenter la connaissance biologique. Cependant elles ne permettent pas de compter les entités consommées ou produites par les processus et par conséquent de représenter les contraintes des modèles. Nous avons donc entrepris la construction d'une nouvelle ontologie.

3 Eléments d'une bio-ontologie pour l'organisation des données et des connaissances

Nous souhaitons donc construire une représentation formelle en nous basant sur les contraintes qui pilotent la construction de modèles mathématiques. Pour ce faire, nous avons montré qu'il doit être possible de compter les molécules consommées ou produites par les processus. Nous proposons tout d'abord un ensemble de concepts et de relations qui vont nous permettre de compter les molécules (section 3.1). Ces concepts et relations nous permettront ensuite de définir de façon formelle un processus biologique (section 3.2).

3.1 Concepts et relations afin de compter les molécules

Nous reprenons dans ce travail l'approche formelle et le concept de processus biologique (nommé *process*) tel qu'il est défini dans les ontologies BiPON/BiPOm. Afin de prendre en compte les contraintes des modèles (la causalité, la conservation de la masse et la compétition pour les ressources), nous utilisons les concepts fréquemment utilisés par les modélisateurs [16]. Nous créons tout d'abord le concept de *pool* qui permet de regrouper toutes les molécules de la même entité chimique. Par exemple toutes les molécules d'eau seront regroupées dans un pool nommé H2O. Ensuite, le pool ayant un volume fini, le nombre de molécules présentes est donné par la *concentration* du pool. Nous décrétons aussi que les processus communiquent uniquement par les pools. Pour ce faire nous imaginons trois opérations (lecture, consommation et production) : (i) un processus peut lire (*reads*) la *concentration* de molécules d'un *pool* et (ii) un processus peut consommer les molécules d'un *pool* et/ou produire les molécules d'un *pool*. Nous représentons ces opérations de consommation/production avec la relation *triggers* et le concept de flux (*flow*).

La figure 2.b reproduit l'exemple de la figure 1.b où le processus P convertit la molécule A en molécule B. Cette figure peut être détaillée ainsi : le processus (*process*) P lit (*reads*) la *concentration* de molécule du *pool* A (flèche grise pointillée). S'il y a suffisamment de molécule (ici une seule molécule est requise), P consomme cette molécule (on dira que P déclenche (*triggers*) un flux (*flow*) de molécule A (première flèche noire)) et produit un flux de molécule B dans le *pool* B (P déclenche un flux de molécule B, deuxième flèche noire).

3.2 Définition formelle du processus biologique

L'ensemble de concepts et de relations décrits ci-dessus nous permet d'affiner la définition de processus biologique proposé par BiPON/BiPOm. Dans ces ontologies le processus est décrit à travers les relations qu'il entretient avec les molécules qui participent à la réaction : un processus a comme entrée (*has_input*) des molécules et comme sortie (*has_output*) des molécules. Nous proposons de faire évoluer cette description. Cette évolution est expliquée à travers les exemples des figures 1 et 2. Tandis que BiPON/BiPOm décrivent le processus P avec comme entrée la molécule A et comme sortie la molécule B, nous déclarons que le processus P lit la *concentration* de molécules A et (s'il y a suffisamment de molécules) déclenche un flux de molécules A et un flux de molécules B. Cette nouvelle description nous permet d'être plus en adéquation avec la contrainte de causalité : l'information donnée par la lecture de la concentration (i.e. le fait qu'il y ait suffisamment de molécules) est la cause du comportement du processus alors que le flux de molécules est considéré comme son effet. Avec ces considérations nous pouvons fournir une définition du processus biologique. Un processus biologique est

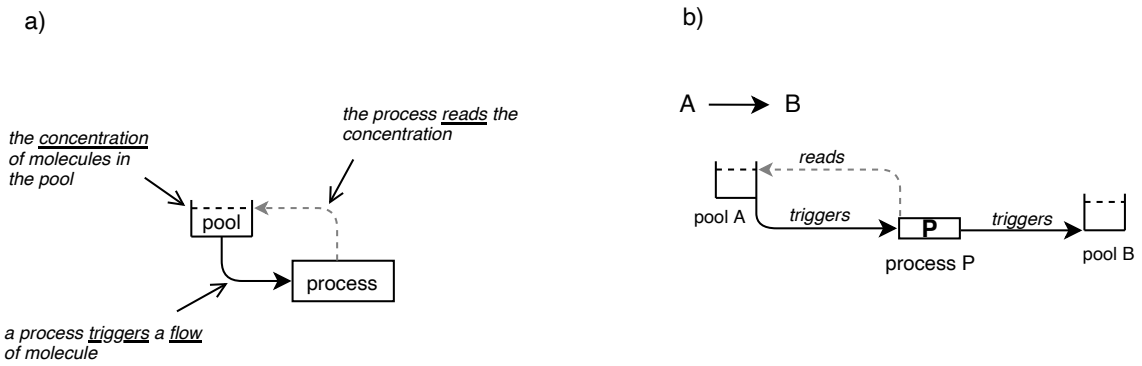


FIGURE 2 – a) Les concepts et relations de la nouvelle ontologie. b) Une réaction biochimique représentée par ces concepts et ces relations.

caractérisé par ses relations avec ses entrées et ses sorties :

$$\mathit{BiologicalProcess} \equiv \exists \mathit{has_input}.Input \sqcap \forall \mathit{has_input}.\mathit{Input} \sqcap \exists has_output.Output \sqcap \forall has_output.Output$$

Une entrée est la concentration lue par un processus :

$$\mathit{Input} \equiv \mathit{Concentration} \sqcap \exists \mathit{is_read_by}.\mathit{BiologicalProcess}$$

Une sortie est le flux de molécules déclenché par le processus :

$$\mathit{Output} \equiv \mathit{Flow} \sqcap \exists \mathit{triggered_by}.\mathit{BiologicalProcess}$$

Il faut noter que la définition du processus est cyclique : il est défini par ses entrées et ses sorties et elles même sont définies par le processus. Ces situations sont communes dans la mise au point d'ontologies. Les cycles peuvent être résolus lors du peuplement en précisant l'ordre selon lequel les individus sont créés.

4 Exemple

Nous illustrons l'utilisation de l'ontologie avec l'exemple d'une réaction catalysée par une enzyme. Cette classe de réaction est représentative d'une large part des processus métaboliques à l'oeuvre dans la cellule bactérienne. (Il faut aussi noter qu'un tiers des gènes de la bactérie sont impliqués dans la synthèse d'enzymes.) Par conséquent, si le processus de catalyse peut être représenté par les concepts et relations décrits dans la section 3.1, nous aurons accompli une première étape dans le processus d'évaluation de l'ontologie. Le modèle chimique qui décrit la catalyse a été proposé par Michaelis et Menten [12]. Ce modèle comprend 2 réactions :

$$E + S \leftrightarrows [ES] \rightarrow E + P$$

Dans la première réaction l'enzyme *E* se lie au substrat *S* pour former le complexe *ES*. Cette réaction est réversible : le complexe *ES* peut se dissocier pour relâcher l'enzyme *E* et le substrat *S*. La deuxième réaction est irréversible : le complexe *ES* se dissocie pour relâcher l'enzyme *E* et le produit *P*.

Afin de représenter ce modèle chimique avec les concepts et relations proposés ci-dessus nous construisons tout d'abord deux processus P1 et P2, pour la première et la seconde réaction. Pour chaque type de molécule nous construisons quatre pools nommés *S*, *E*, *ES* et *P* qui correspondent respectivement au substrat, à l'enzyme, au complexe et au produit. Les processus, les pools et leurs relations sont présentés sur la figure 3.b. Cette figure peut être lue comme suit : P2 lit la *concentration* du *pool* ES et (si la quantité de molécules ES est suffisante) déclenche un flux de molécules E, P et ES. Le processus P1 représente une réaction réversible. Pour la première réaction élémentaire (E+S $\longrightarrow$ [*ES*]) P1 lit la concentration du *pool* E et S et déclenche un flux de E, S et ES. Pour la seconde réaction élémentaire ([*ES*] $\longrightarrow$ E+S) P1 lit la concentration du *pool* ES et déclenche un flux de ES, E et S. Ainsi, dans l'ontologie les deux réactions élémentaires sont agrégées dans un seul processus. Il faut noter que suite à cette agrégation la causalité est toujours respectée. En effet, les sorties du processus P1 (les flux de ES, S et E) sont bien causées par le niveau de concentration des pools ES, S et E.

5 Conclusion et perspectives

Nous avons décrit dans cet article les premières étapes du développement d'une ontologie dédiée à l'organisation des données biologiques. Cette ontologie a été construite en fonction des contraintes qui régissent la construction des modèles mathématiques. L'ensemble de concepts et relations qui en résulte (i) rend possible la représentation des quantités, (ii) a été validé sur un exemple représentatif et (iii) nous a conduits à donner une nouvelle définition formelle du processus biologie. Nous prévoyons maintenant de peupler l'ontologie avec un réseau complet de réactions [5] au format SBML [7]. Lors de cette opération nous pourrons associer des quantités aux concentrations et aux flux. De plus, nous évaluerons la capacité du langage SHACL à exprimer les contraintes des modèles. Ce travail s'inscrit dans un contexte où il s'agit de rendre les ontologies plus expressives avec l'idée de représenter de la connaissance quantitative. Nous pourrions ainsi vérifier la consistance et

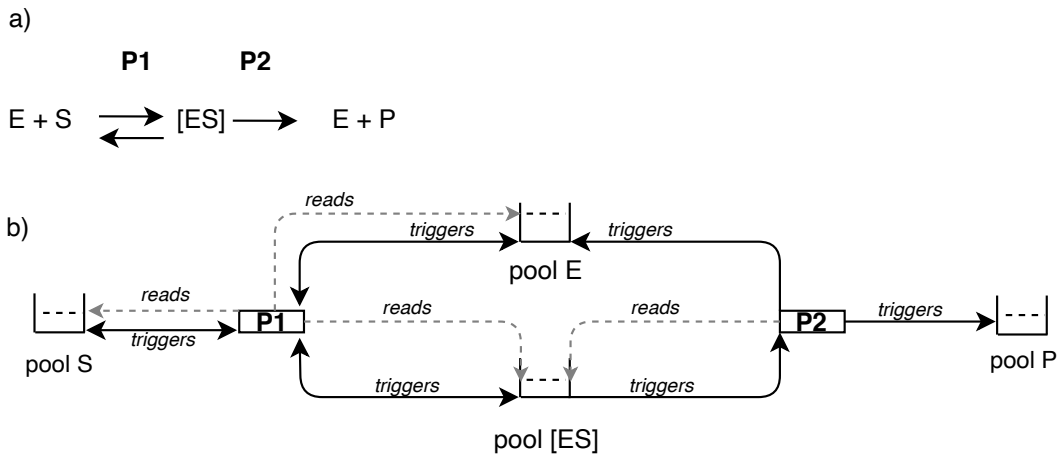


FIGURE 3 – a) Le modèle de catalyse de Michaelis et Menten b) La représentation de ce modèle avec les processus, les pools et les relations.

la validité des données biologiques organisées par cette ontologie.

Références

[1] Gene Ontology Consortium. The gene ontology resource : 20 years and still GOing strong. *Nucleic acids research*, 47(D1) :D330–D338, 2019.

[2] Emek Demir, Michael P Cary, Suzanne Paley, et al. The BioPAX community standard for pathway data sharing. *Nature biotechnology*, 28(9) :935–942, 2010.

[3] Anne Goelzer, Jan Muntel, Victor Chubukov, et al. Quantitative prediction of genome-wide resource allocation in bacteria. *Metabolic engineering*, 32 :232–243, 2015.

[4] Leland H Hartwell, John J Hopfield, Stanislas Leibler, et al. From molecular to modular cell biology. *Nature*, 402(6761) :C47–C52, 1999.

[5] Vincent Henry, Fatiha Saïs, Olivier Inizan, et al. BiPOm : a rule-based ontology to represent and infer molecule knowledge from a biological process-centered viewpoint. *BMC bioinformatics*, 21(1) :1–18, 2020.

[6] Vincent J Henry, Anne Goelzer, Arnaud Ferré, et al. The bacterial interlocked process ONtology (BiPON) : a systemic multi-scale unified representation of biological processes in prokaryotes. *Journal of biomedical semantics*, 8(1) :1–16, 2017.

[7] Michael Hucka, Andrew Finney, Herbert M Sauro, et al. The systems biology markup language (SBML) : a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4) :524–531, 2003.

[8] Olivier Inizan, Vincent Fromion, Anne Goelzer, Fatiha Saïs, and Danai Symeonidou. An ontology to structure biological data : thecontribution of mathematical models. In *Metadata and Semantic Research - 15th International Conference, MTSR 2021*, Communications in Computer and Information Science. Springer, 2021.

[9] Andrew R Joyce and Bernhard Ø Palsson. The model organism as a system : integrating'omics' data sets. *Nature reviews Molecular cell biology*, 7(3) :198–210, 2006.

[10] Jonathan R Karr, Jayodita C Sanghvi, Derek N Macklin, et al. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2) :389–401, 2012.

[11] Evangelina López de Maturana, Lola Alonso, Pablo Alarcón, et al. Challenges in the integration of omics and non-omics data. *Genes*, 10(3) :238, 2019.

[12] Leonor Michaelis, Maud L Menten, et al. Die kinetik der invertinwirkung. *Biochem. z*, 49(333-369) :352, 1913.

[13] Jacques Nicolas. Artificial intelligence and bioinformatics. *A Guided Tour of Artificial Intelligence Research*, pages 209–264, 2020.

[14] Charlotte Ramon, Mattia G Gollub, and Jörg Stelling. Integrating–omics data into genome-scale metabolic network models : principles and challenges. *Essays in biochemistry*, 62(4) :563–574, 2018.

[15] Jason A Reuter, Damek V Spacek, and Michael P Snyder. High-throughput sequencing technologies. *Molecular cell*, 58(4) :586–597, 2015.

[16] Eberhard O Voit. *Computational analysis of biochemical systems : a practical guide for biochemists and molecular biologists*. Cambridge University Press, 2000.

Session 9 : Ontologies et raisonnement pour les systèmes complexes

OWL2YAMS : créer une application CubicWeb à partir d'une ontologie OWL

Fabien Amarger, Nicolas Chauvat, Elodie Thiéblin

Logilab, 104 boulevard Louis-Auguste Blanqui 75013, Paris

prénom.nom@logilab.fr

Résumé

CubicWeb est un cadriciel Web qui facilite le développement de systèmes de gestion de contenus sémantiques (SCMS), lesquels permettent la publication de données par négociation de contenu HTTP et leur administration avec les fonctions attendues d'un CMS classique. Une application CubicWeb est basée sur un schéma YAMS qui définit la modélisation métier des données gérées. OWL est le standard du W3C pour décrire des modèles de données. Nous présentons dans cet article l'outil OWL2YAMS qui permet de créer une application CubicWeb à partir d'une ontologie OWL, puis de charger dans cette application des données décrites en RDF avec les termes de cette ontologie.

Mots-clés

CubicWeb, Web de données liées, négociation de contenu, publication RDF

Abstract

CubicWeb is a semantic framework dealing with data publication, content negociation and data administration, as well as providing Web framework basic functionalities. A CubicWeb application is based on a YAMS schema to model the data. OWL is the W3C standard for representing data model. In this article, we present OWL2YAMS to create a CubicWeb application from an OWL ontology. This tool also enables the loading of data encoded in RDF with the vocabulary of this ontology.

Keywords

CubicWeb, Linked data, content negociation, RDF publishing

1 Introduction

Les données du Web de données liées sont très souvent publiées dans un entrepôt SPARQL ou en utilisant un fichier "dump" contenant tous les triplets RDF. Il est plus rare de pouvoir accéder aux données à travers la négociation de contenu en déréférençant les URI. De plus, nous ne connaissons pas d'interface utilisateur permettant de manipuler des données RDF à la manière d'un CMS (opérations CRUD [1], définitions fines des permissions, rendu graphique, etc.). C'est pour cela que nous présentons CubicWeb, notre cadriciel pour développer des systèmes de gestion de contenu (CMS [2]) sémantiques, accompagné de OWL2YAMS qui permet de créer une application CubicWeb à partir d'une ontologie OWL et d'importer des données en RDF. Il devient ainsi très simple de publier des données RDF et une ontologie OWL pour les intégrer au Web des données liées, en utilisant la négociation de contenu et en bénéficiant de toutes les fonctionnalités et interfaces d'administration que l'on peut attendre d'un CMS sémantique.

2 CubicWeb

CubicWeb est un logiciel libre écrit en Python, dont le développement a commencé en 2001, l'année de publication de l'article fondateur du Web Sémantique [1]. CubicWeb a été conçu pour faciliter le développement et le déploiement d'applications qui reprennent les concepts essentiels du Web Sémantique. Avec CubicWeb il est aisé de gérer et de rendre accessibles des données qui suivent un modèle préalablement défini.

CubicWeb utilise le formalisme YAMS (Yet Another Magic Schema [3] pour représenter le modèle de données de façon explicite. Observons sur la figure 1 que ce schéma YAMS est utilisé pour générer un modèle de données SQL. De cette façon, le schéma explicite de YAMS s'appuie sur la performance et la stabilité du SQL pour son fonctionnement technique.

Le langage RQL (Relation Query Language [4] est utilisé pour exprimer des requêtes en utilisant les noms des classes et des relations du modèle exprimé en YAMS. La requête RQL est compilée en une requête SQL et exécutée sur la base de données relationnelle. Cette approche permet de développer une application Web en ne manipulant que les termes d'un modèle de données explicite et donc de s'abstraire des contraintes techniques du SQL sous-jacentes. L'avantage est que le modèle YAMS et les requêtes RQL peuvent être présentés aux experts métiers et servir de base

1. Create, Retrieve, Update, Delete

2. https://en.wikipedia.org/wiki/Content_management_system

3. https://forge.extranet.logilab.fr/open-source/yams

4. https://forge.extranet.logilab.fr/cubicweb/RQL

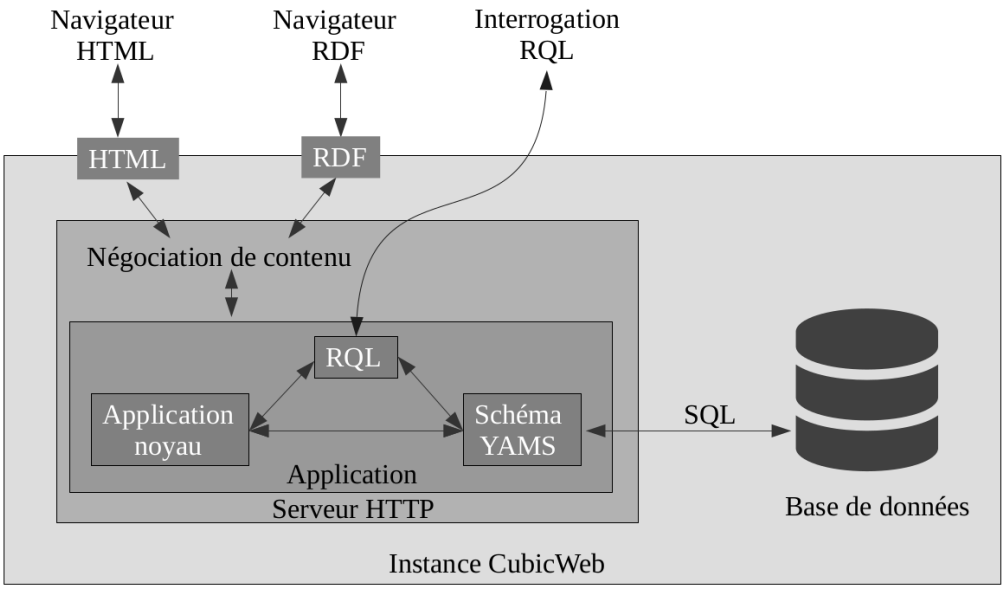


FIGURE 1 – Schéma d'une instance CubicWeb en fonctionnement

de discussion commune, alors que les contraintes techniques du SQL (normalisation, tables de jointure, clef étrangère, etc.) obscurciraient le tableau et rendraient la discussion plus difficile.

Le formalisme YAMS permet la définition de contraintes fortes sur les données (cardinalités, permissions en lecture et/ou en écriture, etc.). Un grand nombre de contraintes métiers peuvent donc être définies directement dans le modèle de données avec la garantie d'être respectée par les requêtes RQL en lecture et en écriture. L'étape de compilation de RQL en SQL combine en effet le contenu de la requête RQL et les contraintes exprimées dans le modèle YAMS. La même requête RQL exécutée par des utilisateurs ayant des permissions différentes ne produit pas toujours le même SQL et ne retourne pas nécessairement le même résultat.

Une fonctionnalité particulièrement appréciable dans CubicWeb est la possibilité d'utiliser une interface graphique automatiquement générée. Cette interface permet d'administrer les données facilement, de gérer les utilisateurs et utilisatrices, d'afficher les données, etc. Cette interface n'est pas obligatoire et il est tout à fait possible de développer sa propre interface grâce à une API permettant de manipuler les données facilement.

Chaque application développée avec CubicWeb prend la forme d'un *cube*, c'est-à-dire d'un composant applicatif réutilisable définissant un modèle, une logique métier et des règles d'affichage. Pour développer une application, il est possible d'inclure d'autres cubes, ce qui signifie qu'avec de l'organisation des applications peuvent être construites par assemblage de composants fonctionnels élémentaires.

Fort de ces fonctionnalités, CubicWeb a pu répondre aux besoins de projets conséquents tels que DataBnf (https ://-data.bnf.fr) [2], FranceArchives (https ://francearchives.fr/) ou DataPOC (https ://datapoc.mnhn.fr/).

Pour une description plus détaillée de CubicWeb, on pourra lire [3].

3 État de l'art

3.1 Application de publication sur le Web de données liées

La question principale sur laquelle nous nous penchons est : "comment construire facilement une application basée sur des données en RDF ?" Le premier pas est la publication des données sur le Web [4]. Nous faisons ici l'état de l'art des systèmes permettant la gestion de ces deux étapes : exposer des ressources RDF sur le web et proposer une interface d'administration pour consulter et/ou gérer (créer, modifier ou supprimer) ces ressources.

Dans cet état de l'art, nous avons considéré les propriétés suivantes comme nécessaires à la gestion des données :

négo. cont. Si l'outil propose de la négociation de contenu sur les données gérées et quelle forme prend cette négociation de contenu (orientée serveur, agent ou transparente, c'est-à-dire par redirection HTTP).

SPARQL Si l'outil propose une API SPARQL pour interroger les données

RDF S'il est possible d'intégrer des données existantes en RDF dans l'outil

perms Si l'outil propose une gestion des permissions sur les données

admin. Si une interface d'administration des données est incluse dans l'outil

visu. Si une interface de visualisation des données (pages html) est incluse

licence La licence de publication de l'outil

publication La date dernière version publiée

Outil	négo. cont.	SPARQL	RDF	perms.	admin.	visu.	licence	publication	inst.
Apache Marmotta [1]	serveur	✓	✓pour représenter ressources LDP	✓	✓	✗	Apache 2	12/06/2018	✓
CarbonLDP [2]	serveur	✓	✗	✓	✓	✗	Propriétaire	4/10/2018	✓Version Gratuite
OntoWiki [3]	✗	✓virtuoso dédié	✓	✗	✓	✓	GPL	4/10/2016	✗Erreur PHP à l'exécution
Open Semantic Framework [4]	?	✓	?	✓	✓	✓	Apache 2	26/02/2015	✗Limité à CentOS 6 et 7, Ubuntu 14.04
PSPS [5]	serveur	✓	✓	✗	✗	✓	MIT License	17/11/2019	✓
Virtuoso [6]	transparent [9]	✓	✓	✓	✗	✓	GPL	22/06/2021	✓Version OpenSource
LinkedDataHub [7]	server (uniquement sur les graphes)	✓	✓	✓	✓	✓	Apache-2.0 License	18/02/2022	✓
CubicWeb [8]	serveur	✗	✓	✓	✓	✓	LGPL	10/03/2022	✓

[1] `https://marmotta.apache.org`
[2] `https://carbonldp.com/`
[3] `https://docs.ontowiki.net`
[4] `http://opensemanticframework.org`
[5] `https://github.com/factsmission/psps`
[6] `https://virtuoso.openlinksw.com`
[7] `https://atomgraph.github.io/LinkedDataHub/`
[8] `https://cubicweb.readthedocs.io/`
[9] `https://datatracker.ietf.org/doc/html/rfc2295#section-4.3`

TABLE 1 – Comparaison d'outils de publication et gestion de données RDF sur le Web

inst. Si nous avons réussi à installer l'outil

Le tableau 1 présente le résultat de cette étude.

D'après les contraintes que nous nous étions fixées, nous pouvons observer que CubicWeb est le seul cadriciel proposant de la négociation de contenu, ainsi que la gestion des permissions et une interface de visualisation et d'administration des données. De plus, CubicWeb est aussi toujours maintenu à l'heure actuelle et publié sous une licence libre LGPL. Nous pouvons remarquer néanmoins qu'il ne permet pas d'interroger les données en SPARQL. Ce manque est en cours d'étude, mais comporte quelques problèmes techniques, principalement liés au fait que RQL est bien moins expressif que SPARQL.

Le projet *LinkedDataHub* semble aussi correspondre à nos critères et semble lui aussi maintenu, mais nous avons remarqué quelques limitations. Tout d'abord, la négociation de contenu ne peut se faire que sur un graphe et non pas sur une ressource particulière. De plus, la spécification de la visualisation d'une ressource ne peut se faire qu'à partir de règles CSS [5], ce qui ne permet pas de répondre à tous les besoins en visualisation, notamment lorsqu'une interaction avec l'utilisateur ou l'utilisatrice est nécessaire.

5. `https://www.w3.org/Style/CSS/Overview.en.html`

3.2 Comparaison OWL/YAMS

Comme détaillé dans la table 2, YAMS est moins expressif que les profils OWL. Il ne couvre pas entièrement le fragment $\mathcal{AL}$ de la logique de description, car il ne permet pas d'exprimer une intersection de concepts. Il s'éloigne également de OWL-Lite par l'absence de transitivité des rôles et leur hiérarchie.

Certains fragments de la logique de description sont toutefois partiellement couverts par YAMS comme la négation, la disjonction de concepts, le "un-de" ou la quantification existentielle typée. Ces expressions sont possibles en YAMS uniquement dans la définition du domaine ou du co-domaine (*range*) d'un rôle.

Comme OWL-Lite, il permet d'exprimer des restrictions de cardinalité pour les valeurs 0 et 1 uniquement [6].

4 OWL2YAMS

CubicWeb permet la négociation de contenu et l'administration de données basées sur un modèle de données contenant des connaissances métiers exprimés en YAMS. Afin de s'intégrer dans l'environnement des standards du W3C, nous avons développé un module de traduction

6. Pour une description détaillée des primitives YAMS-OWL, voir `https://forge.extranet.logilab.fr/open-source/SemWeb/cubicweb_W3C_standard/-/blob/branch/default/yams_owl.csv`

Logique de Description	formule	YAMS	OWL-Lite	OWL-DL	OWL2 Full
$\mathcal{AL}$	C				
$\mathcal{AL}$	$\top$				
$\mathcal{AL}$	$\forall R.C$				
$\mathcal{AL}$	$C_1 \sqcap C_2$				
$\mathcal{F}$					
$\mathcal{U}$	$C_1 \sqcup C_2$				
$\mathcal{C}$	$\neg C$				
$\mathcal{AL}$	R				
$\mathcal{E}$	$\exists R.C$				
$\mathcal{H}$	$R_1 \sqsubseteq R_2$				
$\mathcal{R}^+$	R^+				
$(\mathcal{D})$					
$\mathcal{I}$	R^-				
$\mathcal{O}$	$\{a_1, \ldots, a_n\}$				
$\mathcal{B}$	$\exists R.\{a\}$				
$\mathcal{N}$	$(\geq n\ R)$ ou $(\leq n\ R)$				
$\mathcal{R}$	$R_1 \sqcap R_2$				
$\mathcal{Q}$	$(\geq n\ R.C)$ ou $(\leq n\ R.C)$				

TABLE 2 – Comparaison d'expressivité entre YAMS et les profils OWL (Lite, DL, Full)

OWL2YAMS [7], qui permet de créer une application CubicWeb à partir d'une ontologie exprimée en OWL.

Ce script se base sur une traduction en YAMS d'une ontologie OWL. Comme présenté dans la section 3.2, YAMS est moins expressif que OWL. Tous les axiomes de l'ontologie ne seront donc pas traduits par ce script. Les cases en vert de la colonne YAMS dans la Table 2 ainsi que l'union de concepts dans la définition des domaines et co-domaines des rôles peuvent être traduits par ce script. Les restrictions de cardinalité (limitées à 0 ou 1 dans YAMS) ne sont pas prises en compte pour l'instant.

La figure 2 présente le fonctionnement général de OWL2YAMS. Chaque *owl:Class* ou *rdfs:Class* de l'ontologie est transformée en type d'entité dans le modèle YAMS. Les *owl:DatatypeProperty* sont transformées en attributs en s'assurant que le *rdfs:domain* correspond au type d'entité YAMS et le *rdfs:range* est utilisé avec une correspondance entre les types principaux de *XSD* et les types YAMS. Les *ObjectProperty* sont transformées en relations entre les types d'entités.

Une table de correspondances permet de conserver les URI des éléments de l'ontologie dont sont issus les éléments du schéma YAMS. Cette table est utilisée plus tard pour l'import de données RDF et pour leur traduction en RDF une fois intégrées à CubicWeb (c.f. section 5).

Le script OWL2YAMS crée le schéma YAMS, la structure du cube et l'instance de l'application prête à être lancée et peuplée.

5 Import RDF

Une fois l'application créée, il est possible d'y importer des données tant que ces données sont décrites avec l'ontologie ayant servi à créer cette application. La table de correspondances entre les URI de l'ontologie et les éléments YAMS permet de faire le lien entre données RDF et instance CubicWeb.

La figure 3 schématise l'import de données RDF dans l'application créée à partir de l'ontologie OWL.

De nouvelles URI propres à cette instance CubicWeb sont créées pour chaque individu importé. Les URI originales sont conservées dans la base de données. Ces données sont modifiables par l'utilisateur dans l'interface graphique de CubicWeb.

Les données présentes dans CubicWeb sont téléchargeables en RDF par négociation de contenu via les URI propres à l'instance CubicWeb. La figure 4 reprend l'exemple précédent avec les données d'entrée et celles qui seront renvoyées par la négociation de contenu.

Dans l'exemple, l'URI *http ://<MY-CUBICWEB-BASE-URL>/FoafPerson/123* a été associée à l'URI en entrée *http ://example.org/virginia*. C'est sur *http ://<MY-CUBICWEB-BASE-URL>/FoafPerson/123* que la négociation de contenu sera rendue possible. Un triplet *owl:sameAs* permet de rallier l'URI de l'individu dans l'application et son URI d'origine. Nous avons également pris le parti d'exprimer les éléments du schéma YAMS avec leur propre URI (celle attribuée par l'instance CubicWeb). En effet, les éléments du schéma YAMS ne sont pas strictement égaux à ceux de l'ontologie OWL dont ils découlent, par souci de différence d'expressivité. Pour ne pas perdre la sémantique, nous avons choisi d'exprimer le lien entre ces éléments YAMS et OWL avec une subsomption.

6 Conclusion et perspectives

Nous proposons dans cet article une méthode pour générer une application Web à partir d'une ontologie OWL et

7. https://forge.extranet.logilab.fr/cubicweb/owl2yams/

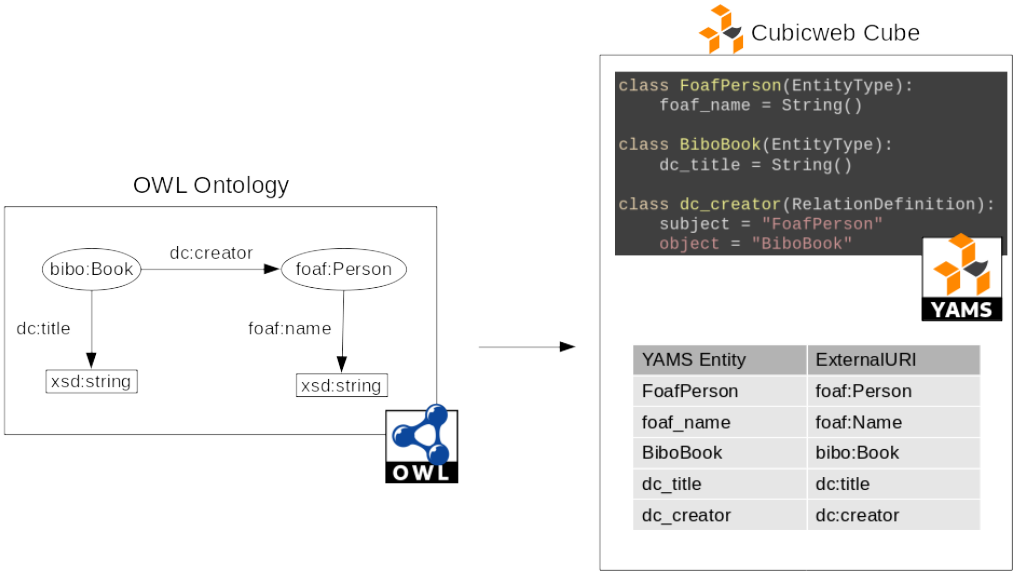


FIGURE 2 – Fonctionnement de OWL2YAMS

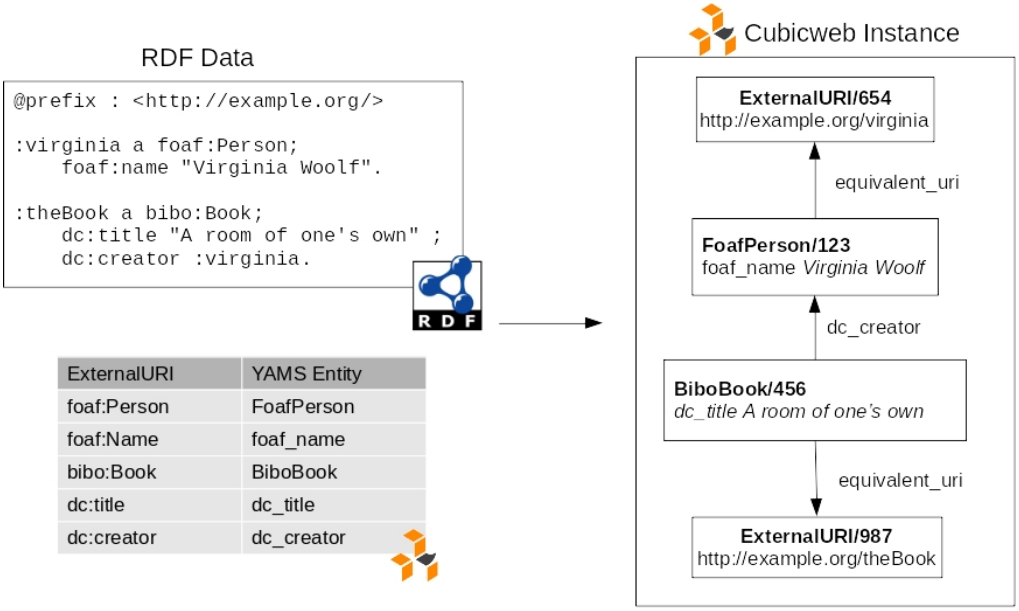


FIGURE 3 – Import de données RDF dans l'application CubicWeb

de données RDF en utilisant OWL2YAMS pour créer un cube et une instance de CubicWeb. De cette manière, il devient très simple de publier des données et l'ontologie associée sur le Web de données liées, en profitant d'un cadriciel complet. CubicWeb propose des fonctionnalités attendues comme la négociation de contenu, la gestion fine des permissions, une interface d'administration, une interface de visualisation et la possibilité de créer sa propre application *front-end* en exploitant les données provenant de CubicWeb.

OWL2YAMS souffre encore de quelques limitations que nous allons lever. À ce stade, certains prédicats de OWL-Lite ne sont pas pris en compte même alors que YAMS permettrait de les exprimer. C'est notamment le cas pour la définition multiple d'attribut (actuellement un attribut ne pourra être défini qu'une seule fois par modèle, par exemple un seul *rdfs:label*), l'héritage multiple de classe ou encore les cardinalités. OWL2YAMS est à ce stade une preuve de concept que nous devrons continuer à développer avant de pouvoir l'utiliser dans des projets en production.

Par ailleurs, dans le but de permettre l'interrogation de CubicWeb avec SPARQL, nous avons étudié les différences d'expressivité entre RQL et SPARQL. Sachant que RQL est transformé en SQL, qui repose sur un paradigme de monde fermé, alors que SPARQL repose sur un paradigme de monde ouvert, la transformation directe de SPARQL en

```
:virginia a foaf:Person;
    foaf:name "Virginia Woolf".

:theBook a bibo:Book;
    dc:title "A room of one's own" ;
    dc:creator :virginia.
```

a) Données RDF en entrée

```
cw:BiboBook rdfs:subClassOf bibo:Book .
cw:FoafPerson rdfs:subClassOf foaf:Person .

cw:dc_author rdfs:subPropertyOf dc:author .
cw:dc_title rdfs:subPropertyOf dc:title .
cw:foaf_name rdfs:subPropertyOf foaf:name .

cw:FoafPerson/123 a cw:FoafPerson ;
    cw:foaf_name "Virginia Woolf" ;
    owl:sameAs <http://example.org/virginia> .

cw:BiboBook/456 a cw:BiboBook ;
    cw:dc_title "A room of one's own" ;
    cw:dc_creator cw:FoafPerson/123 ;
    owl:sameAs <http://example.org/theBook> .
```

b) Données RDF en sortie

FIGURE 4 – Exemple de données RDF en entrée et en sortie de l'application

RQL nous semble difficile. Jusqu'à maintenant nous avons dupliqué les données de CubicWeb dans un entrepôt RDF adjacent chaque fois que l'interrogation en SPARQL était requise. Nous allons continuer à explorer cette question.
Pour finir, notre objectif à moyen terme est de proposer CubicWeb *as a service*, afin que des personnes souhaitant publier des données puissent y parvenir en déposant une ontologie OWL et un graphe RDF dans un formulaire dont la validation déclencherait la création d'une application CubicWeb et son déploiement sur un cluster Kubernetes [8]. Ceci mettrait à la portée de toutes et tous l'intégration de connaissances au Web des données.

7 Bibliography

Références

[1] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 34–43, 2001.

[2] A. Simon, R. Wenz, V. Michel, and A. D. Mascio, "Publishing bibliographic records on the web of data : Opportunities for the bnf (french national library)," in *ESWC*, ser. Lecture Notes in Computer Science, vol. 7882. Springer, 2013, pp. 563–577.

[3] F. Amarger, S. Chabot, N. Chauvat, and E. Thiéblin, "Cubicweb : vers un outil pour des applications clé en main dans le web sémantique," in *31es Journées francophones d'Ingénierie des Connaissances*, 2020.

[4] T. Heath and C. Bizer, "Linked data : Evolving the web into a global data space," *Synthesis lectures on the semantic web : theory and technology*, vol. 1, no. 1, pp. 1–136, 2011.

8. https://kubernetes.io/fr/

Éviter l'échec des systèmes complexes : en construire collectivement une représentation formelle utile

O. Poitou[1], C. Saurel[1]

[1] ONERA Toulouse, 2 avenue Edouard Belin, 31055 Toulouse, France

prenom.nom@onera.fr

Résumé

Cet article présente nos réflexions sur une nécessaire évolution du processus de construction d'un modèle d'un système complexe. Nous identifions le rôle de transcripteur, ses actions et leur insertion dans un processus de construction incrémentale du modèle. Nous proposons des propriétés permettant d'évaluer la qualité de ce modèle au cours du temps. Nous insistons sur la nécessité de capturer la provenance de ses différents éléments, pour répondre à des exigences de traçabilité, de respect de la confidentialité, et d'explicabilité du modèle résultat.

Mots-clés

Ingénierie des modèles, Modélisation collaborative, Modélisation agile, Fusion d'information, Système Complexe

Abstract

We claim that complex system modeling process needs an evolution. We identify the transcriptor role, associated actions and their orchestration in an incremental collective model contruction process. We propose some properties to evaluate model quality at every moment. We promote a precise capture of model elements provenance, as a way to meet traceability, confidentiality and understandability requirements.

Keywords

System Engineering, collaborative modeling, agile modeling, information fusion, complex systems

1 Introduction

La construction collaborative d'un modèle d'un système complexe [1], ou plus précisément d'une spécification formelle avec représentation graphique, est aujourd'hui mal outillée. L'ingénierie système basée modèles (MBSE), discipline récente, n'est déjà plus directement adaptée ni à la complexité actuelle des systèmes, ni à la gestion de la nécessaire collaboration entre les parties prenantes. Les processus et formalismes qu'elle encourage sont parfois trop rigides : certains détenteurs de données expertes ne savent pas les y exprimer. Ils peuvent ne pas être assez précis : les experts peuvent y décrire des informations ne pouvant être exploitées correctement ni par d'autres intervenants ni par des outils d'analyse [8].

Pour éviter ces écueils et mieux outiller la démarche, nous proposons de la formaliser, à travers des rôles, des actions, et des propriétés attendues ou simplement utiles à l'évaluation de la construction du modèle.

La section 2 introduit l'approche et notamment le rôle central de transcripteur. Ses actions et leur insertion dans un processus de construction du modèle du système sont décrites dans la section 4, après la proposition en section 3 de propriétés permettant d'évaluer la qualité de ce modèle au cours du temps. La section 5 positionne ces travaux dans un rapide état de l'art avant que la section 6 ne conclue.

Les exemples utilisés sont issus d'un cas d'étude du domaine de la maintenance aéronautique introduit dans [11].

2 Problématique, approche et rôles

La MBSE propose un cadre à la capture et la restitution des connaissances sur un système, reposant sur l'utilisation de vues standard. Il se veut suffisamment précis pour que le modèle [2] reflète correctement le système modélisé sur des questions d'intérêt : respect de contraintes de sûreté et de sécurité par exemple, indicateurs de performances... Le processus de construction de modèle décrit par ce cadre, appliqué sur un système complexe, devient collaboratif et incrémental puisque les données expertes ne peuvent être obtenues depuis une source unique [12]. Les grandes étapes classiques d'une démarche MBSE gagnent cependant à être assouplies car la synchronisation des différents intervenants peuvent rendre le processus trop contraignant et éloigné de la réalité du terrain. Nous proposons une démarche plus agile capturant l'information lorsqu'elle devient disponible, et la formalisant en un couple modèle/méta-modèle, que nous nommons ici *représentation formelle* [3], en reposant éventuellement sur des interactions, notamment avec le fournisseur de cette information (cette approche est proche du "free modeling" de [6]).

La construction collaborative d'un modèle fait apparaître

1. Une propriété d'un système dit complexe est qu'une seule personne ne peut le connaître entièrement et parfaitement.

2. Le terme *modèle* utilisé ici correspond au terme *Architecture Description* du standard ISO1471/IEEE42010

3. Habituellement, le méta-modèle est d'abord décrit et validé, puis utilisé comme langage pour exprimer le modèle. Dans une démarche plus agile, le modèle et son méta-modèle sont construits en parallèle au fil des intégration des contributions.

différents rôles. Nous nous intéresserons à trois d'entre eux : transcripteur, contributeur et observateur.

Les *contributeurs* sont les dépositaires des connaissances sur le système réel (ou souhaité). Ils vont tenter de le décrire le plus fidèlement possible pour la partie et les aspects qu'ils en connaissent, à travers une ou plusieurs *contributions* (textes, tableaux, schémas... pas nécessairement formels). Chaque contribution apporte des informations nouvelles, des corrections ou des détails supplémentaires.

Les *observateurs* sont les commanditaires ou participent à la maîtrise d'ouvrage. Ils ont des attentes sur le système et souhaitent être informés sur celui-ci dans le respect d'éventuelles règles de confidentialité. Ils réclament des indicateurs pour éclairer leurs prises de décisions, et une information ciblée et présentée selon leur souhait.

Le *transcripteur* est le rôle central de notre approche. Il construit, au fil du temps, la représentation formelle du système en intégrant des contributions émises par les contributeurs, de manière à produire les indicateurs et documentations satisfaisant les observateurs.

3 Propriétés et indicateurs

Pour exploiter de manière fondée le résultat d'une transcription, les observateurs sont en droit d'exiger que le modèle, son méta-modèle, ou même le processus de transcription vérifient certaines propriétés, comme en Génie logiciel ([13]). Nous définissons ci-dessous un tel jeu de propriétés, non exhaustif. Nous suggérons aussi des propriétés souhaitables pour la restitution d'une transcription.

3.1 Propriétés du processus de transcription

Le processus offre la propriété de *traçabilité sur la provenance* quand il permet de fournir la provenance de chaque élément de tout modèle produit. La provenance d'un élément est décrite par les acteurs ayant contribué à son élaboration, et les opérations de fusion d'informations avec leur date et leurs entrées (contributions, ou autres).

La propriété de *transcription sélective* est offerte par le processus s'il permet d'accéder à des versions partielles de la transcription (ie : en ignorant certaines contributions, selon leur horodatage mais aussi leur provenance ou toute autre caractéristique...). Cette propriété peut devenir nécessaire en cas d'exigences de confidentialité entre parties prenantes du système objet de la transcription. La démarche est alors de générer une représentation formelle ne prenant en compte que les contributions auxquelles un destinataire a accès. Ceci garantit qu'aucun élément confidentiel ne peut être retrouvé à partir d'éléments dérivés, que ce soit à travers le méta-modèle, les contraintes/règles retenues etc. [4]

3.2 Propriétés du modèle transcrit

Nous convenons que pour être exploitable par des humains et des outils, un modèle obtenu par transcription doit vérifier les propriétés suivantes.

4. Ce risque existe si l'on se contente de filtrer les éléments confidentiels contenus dans une représentation formelle construite à partir de l'intégralité des contributions

Il doit être *formel* : la transcription doit être décrite dans un langage (ou méta-modèle) interprétable par des outils de calcul ou de raisonnement automatique.

Il doit être *non ambigu* : il ne doit pas permettre deux interprétations différentes de la part des observateurs. La présence d'homonymie est un exemple d'ambiguïté.

Il doit être *complet par rapport aux besoins* formulés par les observateurs. Il doit permettre de satisfaire leurs demandes de documentation par génération de vues pertinentes, et de vérifier l'ensemble des propriétés qui les intéressent, par évaluation de leur expression formelle sur la transcription.

Il doit être *conforme au méta-modèle* actuel obtenu à l'issue de la transcription des contributions disponibles (respect des cardinalités, des types, des contraintes utilisateur).

Il doit être *consensuel* au sein des contributeurs : il doit recueillir au moins une proportion minimale d'approbation de contributeurs, tout en ne dépassant pas une proportion maximale de mise en doute.

Un modèle est *modulaire* s'il peut être structuré en unités indépendantes, réutilisables dans diverses applications.

Ces propriétés concernent le modèle, mais implicitement et conjointement, son méta-modèle. Les propriétés suivantes visent plus directement le méta-modèle.

Le méta-modèle souffre de *superfluité* s'il contient des concepts ou des relations qui ne sont pas instanciés dans le modèle : ces concepts ou relations sont superflus. Une version moins binaire de cette propriété peut être envisagée, limitant les concepts dont le nombre d'instances est sous un certain seuil, ce qui peut montrer un niveau de détail excessif du méta-modèle par rapport au modèle à élaborer.

Il encourage la *concision* si le modèle qu'il permet d'exprimer nécessite d'expliciter peu d'informations déductibles à partir d'autres éléments de ce modèle. Par exemple, indiquer la symétrie d'une relation dans le méta-modèle permet de ne pas devoir systématiquement mentionner les liens dans les deux sens dans le modèle. Notons toutefois qu'indiquer la même symétrie en tant que propriété souhaitée, au lieu de règle, permettra par la suite de vérifier qu'un modèle satisfait cette propriété (au lieu de l'imposer). Cet exemple illustre le compromis à trouver entre automatisation, pour éviter la redondance, et exploitation d'une certaine redondance pour vérifier la cohérence.

3.3 Propriétés sur la restitution

Le processus de transcription vise à satisfaire les besoins des observateurs en produisant régulièrement des observations (cf 4.2) via le processus de restitution.

Une restitution est *complète* si l'ensemble des observations produites répond à l'ensemble des besoins exprimés.

Une restitution est *satisfaisante* si la réponse apportée à chaque besoin est acceptable, par exemple en terme de précision et/ou fiabilité.

Une restitution est *concise* si (une grande proportion parmi) les éléments la composant sont pertinents, i.e. nécessaires et suffisants pour répondre au besoin exprimé.

Une restitution est *juste* si l'information qu'elle véhicule est conforme à celle contenue dans le modèle (le processus de

création d'une représentation n'a pas introduit de contradiction entre l'information présentée et celle du modèle).

Une restitution est *lisible* si, pour chaque observateur, le résultat lui est exprimé dans un langage, et représenté sous une forme, qui lui sont familiers.

4 Actions des différents rôles

Pendant un processus collaboratif et incrémental, les différents intervenants réalisent des actions. Nous définissons ici les actions sur le modèle ou le méta-modèle pour chaque rôle identifié dans la section 2. Des interactions entre ces rôles surviennent en parallèle. Certaines actions peuvent motiver et guider des interactions. A l'inverse, certaines interactions peuvent déclencher les actions ci-dessous.

4.1 Actions du contributeur

Les contributeurs fournissent successivement au transcripteur des informations qui contribuent à la description du système, selon leur expertise et leur point de vue. Ils expriment ces informations dans leur langage favori (par exemple, des schémas de type BPMN) et selon leurs habitudes [5] au travers de trois actions possibles :

- *ajouter* une nouvelle information,
- *corriger* une information existante,
- *raffiner* une information, en la détaillant.

La démarche du transcripteur vise à obtenir un consensus des contributeurs sur la description finale du système : ils devront donc se prononcer sur l'aptitude du modèle transcrit à traduire la vision de chacun sur le système décrit, avec d'autres actions :

- *approuver* le modèle, intégralement ou en partie,
- *mettre en doute* le modèle, ou une partie de celui-ci.

4.2 Actions de l'observateur

Nous distinguons deux types d'*observations* : les *vues* personnalisées et les *indicateurs*.

Les *vues* personnalisées représentent l'ensemble des informations du modèle concernant une attente particulière (un thème, un niveau de détail, une syntaxe concrète). Les actions à réaliser côté transcription sont la sélection d'informations pertinentes et de leur représentation adéquate.

Les *indicateurs* sont les valeurs associées à des expressions. L'action à réaliser côté transcripteur est principalement le calcul. Ces indicateurs peuvent chercher à évaluer le système à travers son modèle, ou bien le modèle en lui-même.

Nous parlerons de *spécification d'observation* pour la demande formulée par un observateur, tandis que nous utiliserons le terme d'*observation* pour la réponse construite par le transcripteur.

Par interaction avec le transcripteur, l'observateur peut :

- *émettre* une spécification d'observation,
- *corriger* une spécification d'observation ou
- *retirer* une spécification d'observation.

5. Ces habitudes proviennent de leur domaine métier, de la charte de leur organisation ainsi que de pratiques personnelles

4.3 Actions du transcripteur

Le transcripteur est en charge de transformer les contributions en un modèle du système d'intérêt, vérifiant les propriétés souhaitées par les contributeurs et observateurs, et permettant d'évaluer les spécifications d'observations émises par les observateurs. Pour cela, le transcripteur procède à plusieurs actions au cours du temps :

- *déchiffrer* une contribution (cf 4.3.1).
- *interpréter* une contribution déchiffrée (cf 4.3.2)
- *consolider* une contribution interprétée (cf 4.3.3)
- *fusionner* un modèle et des contributions (cf 4.3.4)
- *prendre en compte* les spécifications d'observations (cf 4.3.5)
- *évaluer* la représentation formelle (cf 4.3.6)

4.3.1 Déchiffrer

Parfois, une contribution n'est pas directement utilisable et doit être déchiffrée : décrite dans un format exploitable.

Le déchiffrage comprend l'abstraction de certains éléments, son résultat est modifié par les choix d'abstraction réalisés. Dans l'exemple d'une contribution graphique, le déchiffrage peut décrire des textes dans des rectangles, ou conserver les dimensions, et la position exacte de chacun de ces rectangles contenant du texte.

Pour limiter le risque de perdre des éléments porteurs de sémantique, le déchiffrage devra essayer de conserver le plus d'information possible. Notons que cette action est uniquement une action de perception ; les choix d'interprétation seront faits plus tard. Par exemple, le format de représentation des couleurs sera choisi lors de cette action (représentation symbolique, valeurs RVB...) ; mais la question de la sémantique de ces couleurs sera reportée à l'étape d'interprétation : la couleur pourra être utilisée pour participer au typage des éléments, valuer un attribut, ou encore traduire le niveau de d'abstraction, comme dans le NAF (canevas d'architecture de l'OTAN).

Repartir de l'image de la contribution peut être nécessaire, par exemple si une contribution graphique est fournie dans un format de fichier propriétaire (dont le transcripteur n'a pas l'outil correspondant) et que seules les données image sont lisibles. Néanmoins, une représentation utilisable de la contribution peut parfois être directement lue depuis un fichier dans un format ouvert.

4.3.2 Interpréter

Le transcripteur doit générer une représentation formelle de chaque contribution déchiffrée, en vue d'une intégration dans le modèle en cours d'élaboration : on obtient une contribution interprétée. Toutes ses actions doivent conserver les informations de provenance de chaque élément. Pour cela, le transcripteur va, en parallèle :

- faire évoluer un méta-modèle : langage permettant de représenter le contenu des contributions ;
- exprimer dans ce langage chaque contribution, pour en obtenir une version formelle ;
- lier le méta-modèle et la représentation graphique du contributeur dans une feuille de styles.

Par exemple, un carré contenant "Detect" dans une contribution pourra être interprété au niveau méta-modèle par un

concept Activité, au niveau modèle par une entité "Detect" de type Activité, et au niveau style par un lien entre Activité et forme carrée.

Selon le type de contribution, le transcripteur doit :

— enrichir méta-modèle et modèle pour un ajout ou un raffinement,
— réviser méta-modèle et modèle pour une correction.

La complexité de l'opération de révision dépend des cas de figure. Pour les éléments pour lesquels aucune contribution temporellement intermédiaire n'existe, il suffit de remplacer les éléments concernés des contributions à corriger par ceux de la contribution correctrice : on utilisera les mêmes identifiants internes pour ne pas perdre d'éventuels liens avec des éléments non mis à jour. Pour les éléments pour lesquels une contribution intermédiaire existe, un processus de fusion plus avancé doit être mis en place pour décider de la version finale.

4.3.3 Consolider une contribution interprétée

Ces opérations sont à effectuer sur une contribution interprétée, ou un jeu de contributions interprétées d'un même contributeur. Elles facilitent son intégration dans le modèle. Elles sont à placer dans un contexte où les conventions de notation des langages utilisés par les contributeurs ne sont pas toujours respectées, ou les contributions sont faites avec une notation métier. Le transcripteur doit capturer et restituer l'intention du contributeur.

Le transcripteur aura par exemple à identifier des soupçons :

— de lacunes sur des contributions : par exemple, avec des contributions de type BPMN, manque de liens conditionnels entre activités décrites comme déclenchables sur condition, manque de marque de fin globale dans la description d'un processus...
— d'inadéquation du label d'une entité avec son type selon la convention graphique du contributeur,
— de redondances,
— de non conformités par rapport au méta-modèle.

Il devra ensuite :

— proposer des corrections quand un outil permet d'en élaborer, sinon,
— tracer le doute, rattaché à la contribution, dans l'attente de résolution par d'autres contributions.

Le résultat attendu de ces actions sera

— une *contribution consolidée* grâce aux interactions transcripteur/contributeur,
— des informations de provenance concernant les opérations de consolidation,
— une trace regroupant les arguments associés à un soupçon de problème non résolu (pour optimiser ensuite les interactions avec les contributeurs).

4.3.4 Fusionner un modèle et des contributions

Pour intégrer une contribution interprétée supplémentaire au modèle obtenu par intégration d'autres contributions interprétées, tout en assurant les propriétés requises, le transcripteur devra régler les questions qui suivent. Pour cela, il pourra adapter la contribution pour la rendre conforme au méta-modèle courant, ou faire évoluer ce méta-modèle.

Beaucoup de ces questions concernent le label des éléments qui peuplent les contributions. Le *label* d'un élément est son appellation dans une contribution ; il est à distinguer de son *identifiant*. Nous convenons qu'un élément donné n'a qu'un identifiant dans la représentation formelle ; en revanche il peut se voir associer plusieurs labels au fil des contributions, et selon le vocabulaire de chaque contributeur.

Lorsque le transcripteur complète la représentation formelle en cours d'élaboration, en intégrant une contribution consolidée, il doit :

— détecter et aider à résoudre les problèmes soupçonnés d'*homonymie*, sources d'ambiguïté ;
— détecter les occurrences de *presque-homonymie* et aider à les résoudre si elles viennent d'une erreur ;
— détecter les *synonymies*, car elles peuvent générer des incompréhensions entre observateurs ou contributeurs. Les couples (label, utilisateur) doivent être liés à un unique identifiant ; en outre cela permettra d'améliorer la lisibilité des restitutions, en utilisant le vocabulaire adapté à chacun.

Il y a soupçon d'homonymie lorsqu'un label est partagé par des éléments que le transcripteur soupçonne être différents (par exemple, dans une même contribution, plusieurs occurrences d'actions "Defect fix or defer" associées à deux sous-systèmes distincts). On caractérise ce problème selon : le nombre contributions concernées, le nombre de contributeurs, la différence de nature ou de type des éléments (par exemple, label "defect reported" associé à une condition et à un type de message). Selon ces caractéristiques, les arguments permettant au transcripteur de décider qu'il y a ou pas homonymie diffèrent. Le choix du transcripteur (ne rien changer, ou transformer les labels pour enlever l'homonymie) pourra varier entre une décision unilatérale de sa part, et le recours à des interactions avec le(s) contributeur(s) pour lever l'indécision.

Il y a presque-homonymie en cas de labels aux graphies presque identiques (par exemple, "request" et "resquest" ; "request Work order status" et "Work order status request"). Le traitement consiste à déterminer si ces labels désignent des éléments réellement différents (compte-tenu d'une erreur d'un contributeur). Associés à des entités de nature et de types identiques, des liens de labels presque homonymes suggèrent une erreur de graphie. A l'opposé, le transcripteur peut parier qu'il ne s'agit pas d'une erreur de graphie si les entités sont de nature différente (par exemple, relation et concept). Dans de nombreux cas, le transcripteur sera cependant amené à

— élaborer une correction de la transcription pour n'avoir plus qu'un label,
— la faire valider par les contributeurs si le doute reste trop fort pour le transcripteur, ou
— la tracer pour la faire valider plus tard.

Il y a synonymie lorsque plusieurs labels distincts désignent le même élément (par exemple, Aircraft Maintenance Technician et Line Maintenance Operator désignant le même sous-système dans des contributions différentes). Leur étude permet de traduire des éléments du vocabulaire d'un contributeur dans celui d'un autre contributeur.

Le soupçon de synonymie peut intervenir si deux éléments de labels différents reliés aux mêmes éléments apparaissent dans des contributions de sources différentes (hypothèse de vocabulaires différents selon l'organisation ou le domaine métier), ou d'un même contributeur mais espacées dans le temps (hypothèse de changement de vocabulaire du ou des contributeurs au cours du temps). Un même contributeur peut aussi alterner entre deux désignations d'un même élément (hypothèse de synonymie standard). Des actions possibles du transcripteur consistent alors à :

— noter cette hypothèse de synonymie et surveiller l'évolution de sa vraisemblance en fonction de la prise en compte d'autres contributions,
— en cas de synonymie confirmée, ne garder dans le modèle qu'un identifiant, et la liste des labels utilisés par chaque contributeur : cela permettra au transcripteur de générer des vues avec le vocabulaire de leur destinataire.

Il existe d'autres relations. Par exemple [10] définit : l'absence de relation, l'équivalence, l'inclusion et le recouvrement partiel. Nous ne les traitons pas ici.

Ces actions de fusion améliorent la qualité de la représentation formelle. Les informations de provenance conservent les actions effectuées au titre de fusion et leur justification.

4.3.5 Exploiter des spécifications d'observations

Pour assurer la complétude du modèle, le transcripteur doit

— formaliser les spécifications d'observations en exploitant voire enrichissant le méta-modèle,
— les évaluer sur le modèle,
— interagir avec les contributeurs pour compléter les contributions si des données manquent pour évaluer des spécifications d'observations formalisées,
— fournir le résultat de l'évaluation, sous une forme pertinente pour l'observateur. En plus de la spécification d'observation, il s'appuiera sur un profil de l'observateur construit à partir de ses préférences personnelles, celles de l'organisation à laquelle il appartient et celles de son domaine métier [14].

4.3.6 Évaluer la représentation formelle

Pour estimer la qualité de la représentation formelle, le transcripteur peut évaluer les propriétés définies section 3.

Ces évaluations lui permettront de faire remonter les points faibles de la représentation formelle obtenue à partir des contributions considérées. Grâce aux informations de provenance, il sera parfois possible de cibler les contributeurs à solliciter pour améliorer certains points (en leur produisant éventuellement une vue centrée sur ces derniers). Nous estimons que ce mécanisme peut avoir un effet positif très important sur le processus de modélisation collective, en améliorant l'efficience des interactions.

5 Etat de l'art

Dans [2], certaines situations de modélisation sont identifiées comme pouvant se poser lors de la construction agile d'un modèle de système, à partir de contributions diverses. Bien que le positionnement de ces situations y soit fait à un niveau un peu plus abstrait, la démarche est très proche.

La prise en compte de contributions hétérogènes, du point de vue de leur représentation, leur contenu et les concepts utilisés, correspond à l'approche "fédération de modèles" dans [7]. Les auteurs proposent de ne pas construire de méta-modèle commun et de reposer plutôt sur des synchronisations modèle à modèle. Le but est de ne pas imposer à chaque contributeur de reformuler son apport dans un langage qui serait commun à tous, mais moins adapté à l'information qu'il cherche à apporter ou à son domaine métier. Si notre approche tente au contraire de construire un méta-modèle commun, les échanges vers contributeurs et observateurs personnalisent la représentation, afin qu'elle reste pertinente pour le destinataire de l'information.

[5] identifient l'insuffisance du retour sur investissement humain des approches basées modèles comme le principal frein à leur adoption dans le domaine de l'ingénierie logicielle. Nous pensons qu'une observation assez proche peut être faite sur l'ingénierie système, [9] considère qu'il manque principalement des preuves que ce retour est à la hauteur de l'effort consenti. Notre approche va dans le sens de ce que [5] nomment "cognification" de l'ingénierie dirigée par les modèles. Plutôt que de ne reposer que sur le jugement et l'action humaine, nous proposons d'outiller la construction de la représentation formelle pour alléger et guider l'effort humain. Pour cela nous spécifions les actions à mener, et nous identifions certains éléments de décision pour effectuer ou supporter les corrections et ajouts.

Les caractéristiques permettant d'évaluer la qualité de l'information, certaines propriétés du modèle, et certaines idées générales (telle la nécessité d'adapter l'information présentée à un interlocuteur donné) proposées dans cet article, s'inspirent d'éléments utilisés dans le cadre plus général d'évaluation de l'information lors de collaborations ([14]).

L'idée de pouvoir justifier la présence d'un élément d'information par l'historique des opérations ayant contribué à sa genèse n'est pas nouvelle : mais elle donne de la valeur au modèle obtenu par transcription, et aide à convaincre les observateurs du bien-fondé du modèle, et les contributeurs de la prise en compte effective de leurs contributions. L'ontologie PROV-O [6] utilise les concepts d' acteur, d'opération, de donnée, et des relations pour décrire la provenance d'une donnée comme un graphe représentant sa genèse. Par exemple, [4] transforment et résument ensuite de tels graphes pour permettre la réutilisation de données produites en exécution. Nous nous rapprocherons de ces travaux en nous limitant aux besoins de notre projet.

Le contenu de chaque contribution représente un point de vue instantané sur le système à modéliser. Ce contenu est susceptible d'être révisé au vu des contributions venant le compléter ; il peut aussi faire l'objet de plusieurs doutes (sur la terminologie employée, sur des ambiguïtés, etc.) pour le transcripteur qui vise un modèle le moins ambigu possible. Le travail du transcripteur est donc entaché d'incertitude : l'outil destiné à l'aider doit en tenir compte pour pouvoir ajuster au mieux la part d'autonomie de décision du

6. https://www.w3.org/TR/prov-o/

transcripteur et celle des fonctions automatiques. Pour cela, nous pourrons utiliser une technique de représentation des connaissances et de raisonnement avec incertitude ([1]).

6 Conclusion et perspectives

Nous constatons que la complexité des systèmes actuels ne permet plus à un humain d'en réaliser une modélisation exploitable pour en assurer la maîtrise et l'évolution. Nous proposons un processus de transcription d'un jeu de contributions partielles et hétérogènes décrivant un système complexe de manière imparfaite et incomplète. Ce processus, qu'il est possible d'outiller, vise à générer une représentation formelle consensuelle entre les contributeurs. Celle-ci doit répondre aux besoins de ses parties prenantes, dont une gestion sûre de la confidentialité et le calcul d'indicateurs accompagnant leurs décisions sur le système.
Pour outiller ces travaux, nous utilisons un prototype (WEIRD, introduit dans [3]) qui devra être étendu. Ce travail sera enrichi par la prise en compte de divers formats de contribution, le traitement d'autres formes d'ambiguïté sémantique et l'utilisation d'une théorie de l'incertitude.

Remerciements

Ce travail a bénéficié à ses débuts de la participation de Laurence Cholvy (ONERA). Il est en partie financé par l'Agence de l'Innovation de Défense (AID) du Ministère des Armées français (convention de recherche CONCORDE N° 2019 65 0090004707501).

Références

[1] Salem Benferhat, Thierry Denoeux, Didier Dubois, and Henri Prade. Représentations de l'incertitude en intelligence artificielle. In Pierre Marquis, Odile Papini, and Henri Prade, editors, *Panorama de l'Intelligence Artificielle*, volume 1 : Représentation des connaissances et formalisation des raisonnements - Chapitre 3, pages 65–121. Cépaduès Editions, 2014.

[2] Antoine Beugnard, Fabien Dagnat, Sylvain Guerin, and Christophe Guychard. Des situations de modélisation pour évaluer les outils de modélisation. In *INFORSID 2014 : 32ème congrès de l'INFormatique des ORganisations et Systèmes d'Information et de Décision*, pages 181–196, Lyon, France, May 2014.

[3] Pierre Bieber, Frédéric Boniol, Guy Durrieu, Olivier Poitou, Thomas Polacsek, Virginie Wiels, and Ghilaine Martinez. MIMOSA : Towards a model driven certification process. In *8th European Congress on Embedded Real Time Software and Systems (ERTS 2016)*, TOULOUSE, France, January 2016.

[4] Alban Gaignard, Hala Skaf-Molli, and Khalid Belhajjame. Découvrabilité et réutilisation de données produites par des workflows : un cas d'usage en génomique. In Maxime Lefrançois, editor, *Journées Francophones d'Ingénierie des Connaissances (IC) PlateForme Intelligence Artificielle (PFIA'21)*, pages 73–80, Bordeaux, France, June 2021.

[5] Sébastien Gérard, Jordi Cabot, Robert Clarisó, Marco Brambilla, and Sébastien Gerard. Cognifying Model-Driven Software Engineering. In *Software Technologies : Applications and Foundations*, pages 154–160. Springer, January 2018.

[6] Fahad R. Golra, Antoine Beugnard, Fabien Dagnat, Sylvain Guerin, and Christophe Guychard. Using free modeling as an agile method for developing domain specific modeling languages. In *Proceedings of the ACM/IEEE 19th International Conference on Model Driven Engineering Languages and Systems*, MODELS '16, page 24–34, New York, NY, USA, 2016. Association for Computing Machinery.

[7] Fahad Rafique Golra, Antoine Beugnard, Fabien Dagnat, Sylvain Guerin, and Christophe Guychard. Addressing Modularity for Heterogeneous Multi-model Systems using Model Federation. In *MODULARITY 2016 : 15th International Conference on Modularity*, pages 206 – 211, Malaga, Spain, March 2016.

[8] Esther Guerra and Juan de Lara. On the quest for flexible modelling. In *Proceedings of the 21th ACM/IEEE International Conference on Model Driven Engineering Languages and Systems*, MODELS '18, page 23–33, New York, NY, USA, 2018. Association for Computing Machinery.

[9] Kaitlin Henderson and Alejandro Salado. Value and benefits of model-based systems engineering (mbse) : Evidence from the literature. *Systems Engineering*, 24(1) :51–66, January 2021.

[10] Ana Meštrović. Semantic matching using concept lattice. *Concept Discovery in Unstructured Data, CDUD*, 871 :49–58, 01 2012.

[11] Olivier Poitou, Pierre Bieber, Joël Ferreira, and Ludovic Simon. Formal architecture modeling for documenting and assessing Aeronautics Maintenance : A case study. In *ERTS 2018*, 9th European Congress on Embedded Real Time Software and Systems (ERTS 2018), Toulouse, France, January 2018.

[12] A. L. Ramos, J. V. Ferreira, and J. Barceló. Model-Based Systems Engineering : An Emerging Approach for Modern Systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(1) :101–111, January 2012. Conference Name : IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews).

[13] Roxana Saavedra, Luciana C. Ballejos, and M. Ale. Requirements quality evaluation : State of the art and research challenges. In *XIV Simposio Argentino de Ingeniería de Software (ASSE) - JAIIO 42*, pages 240–257, 2013.

[14] Claire Saurel, Olivier Poitou, and Laurence Cholvy. Assessing the usefulness of information in the context of coalition operations. In Éloi Bossé and Galina L. Rogova, editors, *Information Quality in Information Fusion and Decision Making*, pages 135–154. Springer International Publishing, Cham, 2019.

ElvirIA-P : génération d'avis d'expertise pour accompagner les experts en sûreté de fonctionnement des logiciels critiques

K. Cousot[1], T. Sanchez[1], A. Nguyen[1], A. Calpas[2], G. Martinez[2], C. Lopez[1]

[1] Emvista, Cap Oméga, Rond-point Benjamin Franklin, 34960 Montpellier
[2] Agence Innovation Défense (AID), 60 boulevard du Général Martial Valin - CS 21623 - 75509 Paris cedex 15
DGA Techniques Aéronautiques, 47 Rue Saint-Jean, 31130 Balma

{prenom.nom}@emvista.com ; {prenom.nom}@intradef.gouv.fr

Résumé

ElvirIA-P est une preuve de concept ayant pour objectif d'aider l'expert tout au long du processus de certification pour les logiciels embarqués critiques pour la sûreté aéronautique. La solution consiste à extraire des informations d'intérêt pour l'expert dans la masse de documents à sa disposition à partir desquelles un avis d'expertise est généré. Les travaux ont d'abord consisté à représenter les connaissances nécessaires aux experts pour générer un avis d'expertise, puis à extraire les informations des textes pour peupler la base de connaissance et appliquer un raisonnement qui permet de déduire des nouvelles connaissances.

Mots-clés

Preuve de concept, ontologie, raisonnement.

Abstract

ElvirIA-P is a proof of concept aiming at helping the expert throughout the certification process for critical on-board software for aeronautical security. The solution consists in extracting information of interest for the expert in the whole technical documents from which an expert opinion is generated. The work consisted in representing the knowledge necessary for the experts to generate an expert opinion, then in extracting information from the texts to populate the knowledge base and apply reasoning to deduce new knowledge.

Keywords

Proof of concept, ontology, reasoning.

1 Introduction

Au sein de la Direction Générale pour l'Armement (DGA), les experts en sûreté de fonctionnement, du site de DGA Techniques Aéronautiques (DGA TA), spécialisés en logiciel critique doivent faire face à des contraintes fortes : temps limité pour évaluer les logiciels, pénibilité de certaines tâches telles que la lecture de documents volumineux qui nécessite une connaissance *a priori* de formalismes et de langues différents, champs d'expertises vastes, *etc*.

ElvirIA-P est une preuve de concept, financée et encadrée par l'Agence Innovation Défense (AID), ayant pour objectif d'aider l'expert tout au long du processus de certification des logiciels. La solution consiste à extraire des informations d'intérêt pour l'expert dans la masse de documents à sa disposition à partir desquelles un avis d'expertise est généré. Pour l'outil ElvirIA-P, il s'agit d'évaluer les moyens de conformité décrits dans le corpus documentaire pour un logiciel donné. Précisons que les utilisateurs d'ElvirIA-P sont les autorités de certification et non les concepteurs de système. Seul un sous-ensemble du corpus documentaire est transmis à l'autorité et c'est à partir de ce sous-ensemble que l'autorité établit une première évaluation des moyens de conformité proposés par l'appliquant. Les audits et les inspections du corpus documentaire chez l'appliquant vont lui permettre d'établir son avis final. Précisons encore que les auditeurs logiciel (autorité de certification) sont totalement indépendants des développeurs du logiciel. ElvirIA est un outil d'aide pour l'autorité de certification pour l'avis d'expertise.

Dans notre contexte métier qui s'inscrit dans l'aéronautique, le corpus documentaire est constitué des PSAC (*Plan for Software Aspects of Certification*) et SDP (*Software Development Plan*) qui sont les documents décrivant un projet (par exemple le développement d'un logiciel embarqué sur un moteur d'hélicoptère) rédigés par des industriels tels que Thales ou Airbus. Ces documents sont soumis aux experts de la DGA TA qui évaluent comment l'appliquant a pris en compte les recommandations du référentiel métier aéronautique, DO-178, qui énonce un ensemble d'objectifs sur les méthodes et les outils pour le développement d'un logiciel embarqué. Les experts effectuent la mesure de cette conformité par des évaluations des données du cycle de vie du logiciel. L'évaluation se fait généralement en 4 grands audits (planification, développement, vérification et finale). Les documents de certification et les audits permettent alors un audit logiciel et vérifient que les objectifs définis dans le DO-178 sont atteints avant de délivrer un certificat.

Ainsi, pour ElvirIA-P, les textes à analyser sont contenus dans les PSAC et SDP et doivent être confrontés au référentiel métier DO-178. Le développement de ElvirIA-P nécessite :

- une ontologie qui représente les connaissances des experts et les règles métiers associées ;
- un moteur d'extraction d'information faisant appel à des techniques du Traitement Automatique du Langage Naturel pour peupler la base de connaissance ;
- un raisonneur capable d'inférer des connaissances non explicites dans le corpus documentaire ;
- un générateur d'avis d'expertise à partir des données extraites et inférées ;

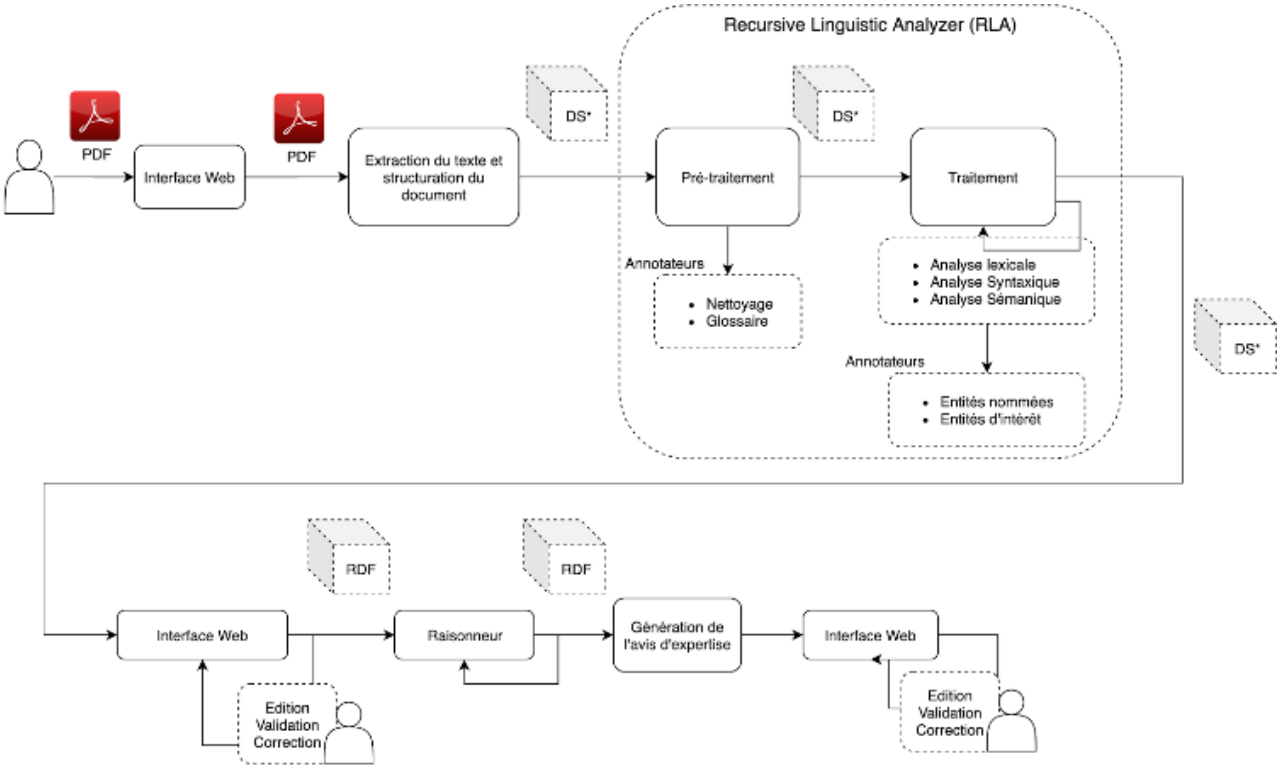


Fig. 1 : Schéma du workflow applicatif

Cet article se focalise sur l'architecture de la solution (cf. section 2), la conception et le développement de l'ontologie.

2 Architecture et flux opérationnel

La conception de l'architecture de ElvirIA-P a été guidée par le flux opérationnel répondant au besoin exprimé par DGA. Le workflow applicatif est décrit de façon séquentielle comme suit (cf. Fig. 1) :

1. L'utilisateur crée un projet sur l'application Web et fournit en entrée un ou plusieurs PSAC/SDP au format PDF ;
2. L'application extrait le contenu textuel et le segmente (sections, paragraphes, …). L'application retourne un Document Structuré, noté DS dans la suite ;
3. Le DS est soumis à un premier traitement de nettoyage (suppression des données trop bruitées, des lignes vides ou avec trop peu de caractères). Cette étape fournit le DS prétraité ;
4. Le DS prétraité est transmis à un analyseur syntaxique et sémantique développé par Emvista pour analyser et extraire les informations pertinentes. L'utilisateur peut voir les éléments identifiés dans les documents via l'interface Web et les corriger si nécessaire avant de les envoyer au moteur de raisonnement ;
5. Une fois les éléments validés par l'utilisateur, le moteur de raisonnement est lancé ;
6. L'avis d'expertise est généré ;
7. L'utilisateur peut encore une fois éditer ou compléter les données générées avant d'enregistrer l'avis.

Compte tenu de ce scénario, une interface Web a été développée avec pour objectif de :

1. permettre à l'utilisateur de modifier des données extraites automatiquement par l'analyseur sémantique avant d'envoyer ces données au raisonneur ;
2. permettre à l'utilisateur de visualiser et d'éditer les avis d'expertises générés automatiquement ;

3 Conception et développement de l'ontologie

Un des premiers objectifs était de développer l'ontologie métier représentant toutes les classes utiles au raisonnement de l'expert

pour générer un avis d'expertise pertinent. Nous avons adopté la méthodologie de Ushold et Gruninger (1996) consistant à identifier les scénarios et les questions de compétences, c'est-à-dire les questions auxquelles notre ontologie devait être en mesure de répondre pour fournir du contenu à l'avis d'expertise. Les scénarios et les questions de compétences ont été élaborés en étroite collaboration entre les ingénieurs de recherche de Emvista et les experts DGA.

Un exemple de scénario est l'identification par un expert des activités traités/non traitées dans les PSAC/SDP, ce qui de fait bloque la délivrance du certificat. Un autre exemple consiste à repérer les techniques mises en œuvre dans les PSAC/SDP et voir si chacune d'entre elles traite l'ensemble des méthodes nécessaires ainsi que les activités relatives à ces méthodes.

Dans le but de déterminer les spécifications de l'ontologie, des questions de compétences ont été identifiées (Gruninger et Fox, 1995), par exemple : « Quelles activités ont été traitées dans les PSAC/SDP ? » ; « Dans quelle section est décrite la méthode A.7-2 ? » ; « Est-ce-que la technique B est partiellement traitée ? ».

Ces questions ont permis d'identifier onze *classes* définies sur deux niveaux, six *object properties* et onze *data properties.*

Trois classes essentielles sont apparues immédiatement et ont

été définies comme suit :

1) *Technique* : une Technique est un ensemble de Méthodes associé à un savoir-faire métier. Les Techniques sont mises en œuvre par l'équipe d'ingénierie développant le système embarquant un/des produits logiciel dans le but d'atteindre la certification. Le nombre de techniques est connu, il en existe huit nommées de A à H. Exemples de techniques : A, « Technique d'aspect système liés au développement logiciel » ; B, « Technique d'organisation des processus du cycle de vie du produit logiciel »)

2) *Méthode* : Une Méthode est un ensemble d'activités qui doivent être correctement traitées afin d'être validée. Le nombre de méthodes est fixé à 74. Exemples de méthodes : A.7-2, « Les résultats des tests sont corrects et les écarts expliqués. ») ; A.4-10, « L'architecture logicielle est compatible avec l'ordinateur cible ».

3) *Activité* : Une activité est un critère d'éligibilité qui agrège des conditions à respecter lors du développement d'un logiciel. Exemple d'activité : Act-5.1.2.a « Les exigences fonctionnelles du système et les exigences d'interface système qui sont allouées au logiciel ne sont pas ambiguës, incohérentes et ne contiennent pas de conditions indéfinies.

D'autres classes permettent de représenter les types d'auditions organisées entre les industriels et les experts DGA (classe Meeting et ses sous-classes, par exemple Familiarization_Meeting et Stages_of_Involvement ») ou encore les niveaux de criticité des logiciels (*Software_Level*).

Les *object properties* identifiées permettent de lier les entités entre elles, avec notamment des relations de méronymie (entre *Software* et *Equipment* par exemple) ou des relations transitives (*is_dependent_of*, entre une Méthode et une Technique et entre une Activité et une Méthode), symétriques (*is_related_to*, entre deux *Engineering_Object*) ou sans propriétés spécifiques (*has_software_level*, entre *Engineering_Object* et *Software_Level*).

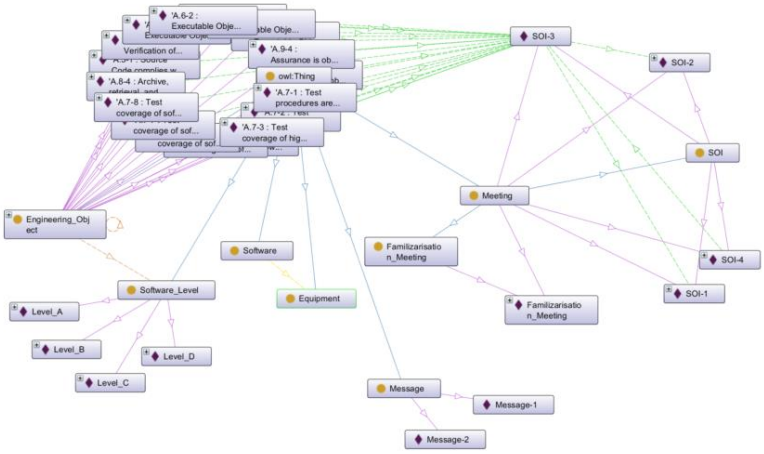


Fig. 2 : Aperçu de l'ontologie développée.

Étant donné que les techniques, les méthodes et les activités sont prédéfinies et n'ont pas vocation à évoluer, elles sont considérées comme des individus nommés : plus de 600 individus sont présents dans la base. Une vue d'ensemble des concepts, relations et individus est donnée en Figure 2. Le développement de l'ontologie a été réalisé avec le logiciel Protégé[1].

[1] https://protege.stanford.edu/

4 Acquisition et développement des règles métiers

Les règles métiers utilisent les concepts de l'ontologie pour générer des informations qui seront restituées dans l'avis d'expertise. Ces règles ont été acquises par le biais de documents textuels de type « référentiels métiers » (précisément la DO-178C qui fixe les conditions de sécurité applicables aux logiciels critiques de l'avionique dans l'aviation commerciale et l'aviation générale) et directement par le biais des experts.
La compréhension du besoin des experts s'est réalisée au cours de nombreux échanges afin de représenter de façon formelle le raisonnement appliqué pour émettre un avis d'expertise. Il en est ressorti que deux familles de règles devaient être prises en compte :

- les règles notées EXP : il s'agit des règles issues du savoir-faire des experts, de leur expérience. Ces règles ne sont explicitées dans aucun document textuel et ont donc été formalisées au cours des discussions.
- les règles REF : il s'agit de règles explicitement décrites dans les référentiels des experts. La difficulté de ces règles est qu'elles sont exprimées en langage naturel et sont difficilement interprétables, y compris par les experts à cause de leur caractère subjectif. Exemple de règle issue de la DO-178C : « *IF a different compiler or different set of compiler options are used, resulting in different object code, THEN the results from a previous software verification process activity using the object code may not be valid and must not be used for the new application. In this case, previous test results may no longer be valid for the structural coverage criteria of the new application. Similarly, compiler assumptions about optimization may not be valid. Therefore, some software verification process activities have to be done again.*".

Étant donné la complexité des règles REF, chacune d'entre elles est représentée par un individu nommé et est liée à l'activité correspondante de sorte à demander à l'expert d'appliquer cette règle le cas échéant via l'avis d'expertise. À noter que les identifiants des paragraphes et de pages de la DO_178C sont également associés à chaque règle, et font partie intégrante de l'ontologie. Leur traitement automatique demeure une perspective à ce travail.

D'une part, les règles REF ont été intégrées dans l'ontologie sous forme textuelle (*cf.* Fig. 1) et non sous forme de règles prises en compte dans le raisonnement pour les raisons évoquées dans la section précédente. Néanmoins, les relations d'association transitives entre la Règle et les Techniques et Méthodes sont prises en compte dans le raisonnement.

D'autre part, les règles EXP ont été développées dans l'ontologie en logique de descriptions avec le langage de programmation SWRL. Ces règles sont donc prises en compte dans le raisonnement d'ElvirIA-P, contrairement aux règles REF. Exemples de règles EXP : Si une activité n'est pas traitée alors la méthode liée à cette activité n'est pas validée (cf. (1) Fig. 4) ; si deux logiciels de niveau de criticité différents sont détectés alors le message 1 doit être retourné (cf. (2) Fig. 4).

```
1) ontology:needs(?m, ?a) ^ ontology:Activity(?a) ^ ontology:is_validated(?a, false) ^ ontology:Method(?m) -> ontology:is_validated(?m, false)

2) ontology:Software(?s) ^ ontology:has_sofware_level(?s, ?dal1) ^ ontology:has_sofware_level(?s, ?dal2) ^ differentFrom(?dal1, ?dal2) -> ontology:is_validated(autogen1:Message-1, true)
```

Fig. 4 : Exemples de règles métiers développées en SWRL

5 Acquisition des données et prétraitement

Après un traitement consistant à sélectionner les documents PSAC/SDP en anglais, au format PDF et non scannés, le corpus d'étude contient 73 fichiers.

La solution Apache PDFBox a été adoptée pour extraire le texte des PDF ; celle-ci prend en entrée un fichier PDF et fournit le texte brut correspondant ainsi que le texte formaté avec HTML. La sortie HTML n'a pas été retenue dans la suite du traitement car celle-ci n'est pas assez fiable (problème fréquent avec les balises <p>). Un module a été développé pour fournir la structure qui contient le texte brut ainsi que son découpage en pages. Celui-ci est composé du Segmenter et du Cleaner décrit ci-après.

Segmenter. Le Segmenter a pour objectif d'identifier des segments textuels particuliers au sein du document. Certains segments seront supprimés par la suite afin de limiter le bruit tandis que d'autre pourront être exploités. C'est d'abord la table des matières qui est recherchée à l'aide de patrons, puis, grâce à l'information qu'elle contient, la structure du document en sections est reconstituée.
Le Segmenter détecte ensuite les en-têtes et les pieds de page : un score de similarité est calculé avec les pages avoisinantes pour mettre en évidence les éléments fortement redondants en début et fin de page. Chacun des segments est ensuite envoyé au Cleaner.

Cleaner. L'objectif du Cleaner est de supprimer le bruit dans le texte extrait. Le bruit trouve son origine dans des défauts du traitement OCR ainsi que dans l'extraction du texte des tableaux et des figures qui ne se présente pas sous la forme d'une phrase analysable par des outils classiques de TALN. Deux méthodes ont été envisagées, l'une faisant partie de la famille d'approche "symbolique" et l'autre faisant partie de la famille d'approche "par apprentissage". La première s'est avérée tout à fait pertinente ce qui a impliqué que la deuxième option a été abandonnée, d'autant plus que celle-ci aurait

nécessité un grand volume de données annotées que nous n'avions pas.

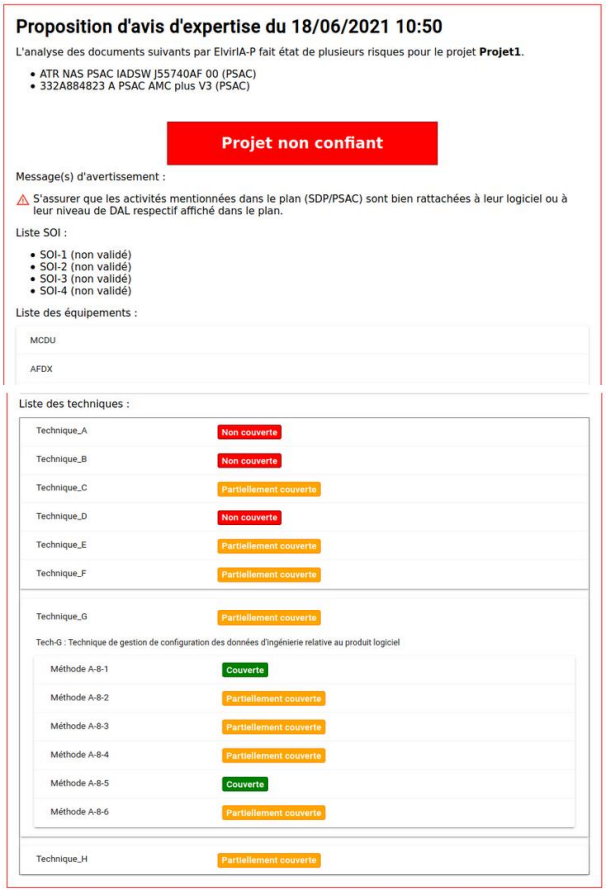


Fig. 4 : Exemple d'une page d'une proposition d'avis d'expertise générée par ElvirIA-P

Une fois que le texte est segmenté et nettoyé, il est envoyé au module d'extraction d'information pour peupler la base de connaissances.

6 Peuplement de la base de connaissance

Étant donné l'ontologie développée, le peuplement de la base de connaissance s'opère *via* deux tâches :

- L'extraction des entités nommées qui permettent de définir le contexte de l'avis d'expertise (numéro de version du document, nom des logiciels concernés, *etc.*) ;
- L'identification des activités traitées à partir desquelles des inférences s'opèrent et qui constituent le fond de l'avis d'expertise ;

La reconnaissance d'entités nommées consiste en un token-level classifier (Devlin et al., 2018) avec une couche de classification après les encodeurs des modèles transformers (Vaswani et al., 2017). Cette couche est une couche dense qui prend en entrée une représentation du "First Sub-Token" (Devlin et al., 2018) et qui fournit une classe (personne, localisation ou organisation) pour chaque mot donné au modèle.

Une surcouche symbolique permet d'identifier les entités nommées pour lesquelles trop peu de données sont annotées (par exemple les noms de logiciels et des équipements).

L'identification des activités est effectuée par la détection de descripteurs spécifiques à une activité (fournis par DGA) qui prennent la forme de mots clés et de triplets exprimés en langage naturel avec utilisation d'opérateurs logiques AND et OR. Concernant les mots-clés, une correspondance est recherchée entre les mots-clés fournis et les mots dans le texte qui sont soumis à une phase de lemmatisation afin de considérer

```
<owl:NamedIndividual rdf:about="https://www.defense.gouv.fr/dga/elviriap/ontology/activity#Act-5.3.2.b">
    <rdf:type rdf:resource="https://www.defense.gouv.fr/dga/elviriap/ontology#Activity"/>
    <rdf:type rdf:resource="https://www.defense.gouv.fr/dga/elviriap/ontology#Engineering_Object"/>
    <is_validated rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean">true</is_validated>
    <rdfs:comment xml:lang="en">The &quot;Source Code&quot; conforms to the &quot;Software Code Standards&quot;.</rdfs:comment>
</owl:NamedIndividual>
```

Fig. 3 : Exemple de triplets RDF relatifs à une activité détectée dans un segment textuel.

La méthode symbolique s'appuie sur des hypothèses de la langue au niveau lexical ; par exemple il est fortement probable qu'une phrase contienne un déterminant ou une majuscule en début de phrase ou encore un point en fin de phrase. Au contraire, la présence soutenue de caractères spéciaux dans une chaîne de caractères tend à montrer que celle-ci n'est pas une phrase bien formée et peut-être le résidu d'un tableau ou d'un schéma, par exemple. Des travaux spécifiques à l'exploitation des figures et des tableaux devraient être menés, par exemple en utilisant Dagobah (Liu et Troncy, 2019). Dans ce module, il n'est pas envisageable d'utiliser une analyse plus profonde du texte (syntaxe, sémantique) puisqu'il s'agit ici justement de préparer le texte à une telle analyse.

les formes fléchies. Exemple de mots-clés pour l'activité Meth-2.1 :

"software components" OR "several software components" / "components" OR "several components" OR "partitioning" / "partition" OR "software partition"

Concernant les relations, une analyse syntaxique du texte est réalisée afin d'identifier les verbes pour lesquels une correspondance est recherchée avec le prédicat du triplet fourni. Une fois le prédicat identifié, son sujet et son objet sont recherchés dans la phrase via les relations syntaxiques « sujet » et « objet ». Cette méthode permet d'identifier des triplets dont

les éléments sont distants dans les textes. Exemple de triplets pour l'activité Act-5.3.2.a :

"source code implement low-level requirement" AND "source code comply with software architecture" OR "source code implement requirement" AND "source code comply with software architecture"

Afin d'augmenter la couverture du traitement, la détection de descripteurs spécifiques s'appuie également sur un niveau sémantique via l'utilisation de synonymes pour les sujets et les objets et de classes sémantiques de VerbNet pour les verbes (Schuler, 2005).

Lorsqu'une information relative à une activité est détectée, ladite activité reçoit la valeur *true*. Un exemple de RDF représentant une activité détectée est donné en Fig. 3. Les activités traitées déclenchent des règles en cascade jusqu'à déterminer si les techniques sont traitées et si les auditions peuvent être organisés.

Nous n'abordons pas ici la génération de l'avis d'expertise à proprement dite mais présentons un exemple (cf. Fig. 4) généré à partir des données structurées et présentes dans la base de connaissance. En quelques clics, l'expert peut accéder à la source (ou aux sources) ayant généré chaque élément présent dans l'avis. Il accède également aux règles EXP impliquées dans ladite génération. Dans le concept ElvrIA, l'objectif est d'évaluer comment une IA peut aider l'auditeur à construire son avis d'expertise.

7 Conclusion

Dans cet article, nous avons présenté ElvirIA-P, une preuve de concept ayant pour objectif d'aider l'expert tout au long du processus de certification pour les logiciels embarqués critiques pour la sûreté aéronautique. Les travaux ont d'abord consisté à représenter les connaissances nécessaires aux experts pour capitaliser, structurer et générer un avis d'expertise, puis à extraire les informations des textes pour peupler la base de connaissances et appliquer un raisonnement qui permet de déduire des nouvelles connaissances.

Des tests du raisonneur ont été réalisés avec différents algorithmes de raisonnement tels que Pellet et Fact++. Ils démontrent que l'ontologie est consistante ; aucun problème d'incompatibilité des règles n'a été détecté.

Les requêtes SPARQL mises en œuvre pour répondre au besoin de la génération d'avis pour les experts de la DGA ont toutes retournées les résultats attendus et permettent effectivement de faire remonter les informations jusqu'à l'interface utilisateur. Les inférences obtenues sont bien celles attendues. Celles-ci sont réalisées en quelques millisecondes.

La preuve de concept n'est cependant pas industrialisable en l'état. La preuve de concept nécessite des études et des travaux de développement supplémentaires pour pouvoir raisonnablement envisagé de la mettre au service d'un expert pour l'accompagner dans son travail. Le projet ElvirIA a montré que les techniques IA peuvent assister dans une certaine mesure le travail de l'expert. Les échanges entre les auditeurs DO-178 et Emvista ont permis de mieux appréhender les possibilités et les limitations des technologies IA disponibles sur le marché sur le traitement du langage. Ils ont mis en évidence que la modélisation de notre métier sur les caractéristiques des technologies du logiciel était très complexe pour l'apprentissage d'une machine contrairement à un référentiel de développement logiciel.

L'analyseur utilisé étant un outil d'analyse de texte bien formée, il a été décidé de se focaliser prioritairement sur les phrases nominales et verbales. Les images, tableaux et listes contiennent néanmoins des informations qui pourront s'avérer indispensables pour la génération d'avis d'expertise et pourront faire l'objet d'un traitement prioritaire dans les prochains travaux, par exemple en utilisant un système d'annotation sémantique de données tabulaires tel que DAGOBAH [7].

8 Références

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

[3] M. Gruninger et M. S. Fox, *Methodology for the Design and Evaluation of Ontologies*, 1995.

[4] Liu, J., & Troncy, R. (2019). DAGOBAH: An End-to-End Context-Free Tabular Data Semantic Annotation System.

[5] Schuler, K. K. (2005). VerbNet: A broad-coverage, comprehensive verb lexicon.

[6] M. Uschold et M. Gruninger, Ontologies: Principles, methods and applications. *The knowledge engineering review*, 11(02), pp. 93-136, 1996.

[7] Chabot, Y., Labbé, T., Liu, J., & Troncy, R. (2020). DAGOBAH: Un système d'annotation sémantique de données tabulaires indépendant du contexte. In *31es Journées francophones d'Ingénierie des Connaissances (Proceedings of the 31st French Knowledge Engineering Conference)*, Angers, France, June 29 - July 3, 2020, pp. 120–132.

Session 10 : TALN, ontologie et graphe de connaissances

Apport des ontologies pour le calcul de la similarité sémantique au sein d'un système de recommandation

LE Ngoc Luyen[1,2], Marie-Hélène ABEL[1], Philippe GOUSPILLOU[2]

[1] Université de technologie de Compiègne, CNRS, Heudiasyc (Heuristics and Diagnosis of Complex Systems), CS 60319 - 60203 Compiègne Cedex, France

[2] Vivocaz, 8 B Rue de la Gare, 002200, Mercin-et-Vaux, France

Résumé

La mesure de la parenté ou ressemblance sémantique entre les termes, les mots, ou les données textuelles joue un rôle important dans différentes applications telles que l'acquisition de connaissances, les systèmes de recommandation, et le traitement du langage naturel. Au cours des dernières années, de nombreuses ontologies ont été développées et utilisées pour structurer les connaissances au sein des systèmes d'information. Le calcul de similarité sémantique à partir d'ontologie s'est développé et selon le contexte est complété par d'autres méthodes de calcul de similarité. Dans cet article, nous proposons et appliquons une approche pour le calcul de la similarité sémantique basée sur l'ontologie au sein d'un système de recommandation.

Mots-clés

Similarité sémantique, Ontologie, Système de Recommandation, Plongement de mots

Abstract

Measurement of the semantic relatedness or likeness between terms, words, or text data plays an important role in different applications dealing with textual data such as knowledge acquisition, recommender system, and natural language processing. Over the past few years, many ontologies have been developed and used as a form of structured representation of knowledge bases for information systems. The calculation of semantic similarity from ontology has developed and depending on the context is complemented by other similarity calculation methods. In this paper, we propose and carry on an approach for the calculation of ontology-based semantic similarity using in the context of a recommender system.

Keywords

Semantic Similarity, Ontology, Recommender System

1 Introduction

Avec le développement d'Internet et du World Wide Web, les sites Web ou applications e-commerce contiennent des de données textuelles structurées, semi-structurées ou non structurées qui ne cessent d'augmenter. La recherche d'informations sur ces sources de données permet d'améliorer certaines tâches telles que la recherche, le classement. Plus précisément, le calcul de la similarité sémantique montre à quel point deux concepts, deux termes ou deux entités sont proches, sur la base de la comparaison des liens taxonomiques et des propriétés sémantiques [32].

En structurant et en organisant un ensemble de termes ou de concepts au sein d'un domaine de manière hiérarchique et en modélisant les relations entre ces ensembles de termes ou de concepts à l'aide d'un descripteur relationnel, une ontologie permet de spécifier un vocabulaire conceptuel standard pour représenter les entités du domaine [30]. Diverses applications utilisant des ontologies décrivent des termes, des entités et quantifient les relations entre eux [29, 18]. Ces dernières années, l'utilisation d'ontologies est devenue plus populaire dans les systèmes de recommandation [15, 27]. Ainsi, le calcul de similarité sémantique basé sur l'ontologie permet d'améliorer la précision des tâches d'appariement, de recherche et de classement sur des éléments ou des profils d'utilisateurs.

Une ontologie peut être représentée selon différents modèles : (1) Le modèle de représentation en triplet définit une ontologie comme un ensemble de triplets $\langle sujet, prédicat, objet \rangle$ où la relation entre le sujet et l'objet est exprimée par le prédicat. Le sujet est une ressource [1], le prédicat est une propriété d'une ressource, et l'objet identifie la valeur de la propriété de la ressource. L'objet d'un triplet peut contenir une autre ressource ou un littéral. (2) Le modèle de représentation graphique considère qu'une ontologie est un graphe orienté où les nœuds représentent les ressources ou les littéraux tandis que les arcs représentent les propriétés nommées. (3) Le modèle de représentation orienté objet définit une ontologie comme un ensemble d'objets, dans lequel les objets correspondent aux ressources et les variables d'instance de l'objet correspondent aux propriétés de ressources [8].

En considérant une ontologie comme un ensemble de triplets, les approches courantes de calcul de similarité sé-

1. Une ressource peut être une classe, un instance, un concept, un nombre, un chaîne de caractères [8]

mantique basées sur une ontologie présentent deux points faibles. Le premier point faible concerne la mesure de similarité qui se calcule soit entre objets, soit entre objets et prédicats [32, 21]. Le calcul basé sur les objets n'utilise pas les informations du sujet, alors qu'elles peuvent contenir des informations contextuelles du triplet intéressantes pour la comparaison. Le second point faible concerne la distinction du type des objets : textuels ou numériques [24]. Le calcul de similarité entre des objets numériques consiste en un simple calcul arithmétique. Le calcul de similarité entre des objets textuels est basé sur la fréquence des mots composant les objets textuels à comparer. Ce calcul ne tient pas compte de la dépendance sémantique entre ces mots. Cette dernière peut être une richesse pour la comparaison. Dans le cadre de nos travaux, nous visons le traitement de ces deux points faibles afin de définir un calcul de similarité sémantique plus précis au sein d'un système de recommandation.
Le reste de cet article est organisé comme suit. Tout d'abord, la section 2 présente des travaux de la littérature sur lesquels s'appuie notre approche. La section 3 présente nos contributions principales sur la construction du système de recommandation exploitant la mesure de similarité entre des ensembles de triplets. Avant de conclure, nous testons nos travaux dans la section 4 à partir d'un cas expérimental traitant de l' achat/vente de véhicules d'occasion. Enfin, nous concluons et présentons les perspectives.

2 Travaux de la littérature

2.1 Apport des ontologies

Dans le contexte du partage des connaissances, une ontologie est une description formelle et explicite des connaissances partagées qui consiste en un ensemble de concepts dans un domaine et les relations entre ces concepts [13]. L'utilisation des ontologies facilite le partage et la réutilisation des connaissances entre les personnes et les applications largement diffusées. L'usage des ontologies permet [2] :

- L'organisation des données : une ontologie est construite sur la base des structures naturelles de l'information en permettant de visualiser les concepts et leurs relations.
- L'amélioration de la recherche : au lieu de rechercher par mot-clé, la recherche sur les ontologies peut renvoyer des synonymes à partir des termes de la requête.
- L'intégration de données issues de différentes sources, différents langages.

Fondamentalement, une ontologie peut être représentée par le langage OWL qui permet de contraindre les faits RDF dans un domaine particulier. Un fait RDF est défini par un triplet qui est un ensemble de trois composants : un sujet, un prédicat et un objet. Intuitivement, un triplet $\langle sujet, prédicat, objet \rangle$ exprime qu'un sujet donné a une valeur donnée pour une propriété donnée[2, 23]. Une ontologie représentée en OWL possède un mécanisme d'inférence ou de raisonnement permettant de déduire les connaissances supplémentaires.
La similarité sémantique basée sur l'ontologie fait référence à la proximité de deux termes [2] au sein d'une ontologie donnée. La distance entre deux termes est une représentation vectorielle numérique de la distance entre deux termes l'un de l'autre [20]. Cela permet d'utiliser l'ontologie pour rechercher efficacement des éléments liés ou pour identifier des associations entre des termes.
L'utilisation des ontologies comme une base de connaissance devient de plus en plus populaire dans les tâches de modélisation, d'inférence des nouvelles connaissances, ou de calcul de similarité pour des systèmes de recommandation [11]. Dans la section suivante, nous rappelons les notions de base des systèmes de recommandation et précisons le rôle que peut y jouer une ontologie notamment dans certains domaines.

2.2 Système de recommandation basé sur les ontologies

Le système de recommandation (SdR) est conventionnellement défini comme une application qui tente de recommander les éléments les plus pertinents aux utilisateurs en raisonnant ou en prédisant les préférences de l'utilisateur dans un élément en fonction d'informations connexes sur les utilisateurs, les éléments, et les intéractions entre les éléments et les utilisateurs [22, 19]. En général, les techniques de recommandation peuvent être classées selon 6 principales approches : les SdRs basés sur les données démographiques, les SdRs basés sur le contenu, les SdRs basés sur le filtrage collaboratif, les SdRs basés sur la connaissance, les SdRs sensibles au contexte, et les SdRs hybrides.
Dans plusieurs domaines tels que les services financiers, les produits de luxe coûteux, l'immobilier ou les automobiles, les articles sont rarement achetés et les évaluations des utilisateurs ne sont souvent pas disponibles. De plus, la description des articles peut être complexe et il est difficile d'obtenir un ensemble raisonnable de notes reflétant l'historique des utilisateurs sur un article similaire. Par conséquent, les SdRs basés sur les données démographiques, sur le contenu, et sur le filtrage collaboratif ne sont généralement pas bien adaptés aux domaines dans lesquels les éléments possèdent les caractéristiques mentionnées. Des systèmes de recommandation basés sur les connaissances représentées au moyen d'ontologies sont alors proposés pour relever ces défis en sollicitant explicitement les besoins des utilisateurs pour ces éléments et une connaissance approfondie du domaine sous-jacent pour les mesures de similarité et le calcul des prédictions [17].
Pour améliorer la qualité de la recommandation, les calculs de similarité entre éléments ou le profil utilisateur dans un système de recommandation joue un rôle très important. Ils permettent d'établir une liste de recommandations tenant compte des préférences des utilisateurs obtenues suite aux déclarations des utilisateurs ou bien de leurs interactions. Nous détaillons dans la section suivante les mesures de similarité sémantique entre les éléments au sein d'un système de recommandation.

2. Une terme est utilisé pour exprimer un concept, un sujet, un prédicat, un objet, ou un ensemble de triplets

2.3 Mesure de similarité sémantique

Les avantages de l'utilisation des ontologies consistent en la réutilisation de la base de connaissances dans divers domaines, la traçabilité et la capacité d'utiliser le calcul et l'application à une échelle complexe et à grande échelle [26]. En fonction de la structure du contexte applicatif et de son modèle de représentation des connaissances, différentes mesures de similarité ont été proposées. En général, ces approches peuvent être classées selon quatre stratégies principales [32, 24] : (1) basée sur le chemin, (2) basée sur les caractéristiques, (3) basée sur le contenu de l'information, et (4) la stratégie hybride qui inclut des combinaisons des trois stratégies de base.

En mesurant la similarité sémantique basée sur le chemin, les ontologies peuvent être considérées comme un graphe orienté avec des nœuds et des liens, dans lequel les classes ou les instances sont interconnectées principalement au moyen de relations d'hyperonyme et d'homonyme où l'information est structurée de manière hiérarchique en utilisant la relation 'est-un' [24]. Ainsi, les similarités sémantiques sont calculées en fonction de la distance entre deux classes ou instances. De cette manière, plus le chemin est long, plus les deux classes ou instances sont sémantiquement différentes [32]. Le principal avantage de cette stratégie est la simplicité car elle nécessite un faible coût de calcul basé sur le modèle de graphe et ne nécessite pas les informations détaillées de chaque classe et instance [21]. Néanmoins, le principal inconvénient de cette stratégie concerne le degré de complétude, d'homogénéité, de couverture et de granularité des relations définies dans l'ontologie [32].

Lors de la mesure des similarités sémantiques basées sur les caractéristiques, les classes et les instances dans les ontologies sont représentées comme un ensemble de caractéristiques ontologiques [32, 24]. Les points communs entre les classes et les instances sont calculés en fonction de leur ensemble de caractéristiques ontologiques. De cette manière, l'augmentation de la différence de deux classes ou instances dépend de l'augmentation de nombreuses propriétés partagées et de la diminution des propriétés non-partagées entre elles [34]. L'évaluation de la similarité peut être réalisée en utilisant plusieurs coefficients sur les ensembles de propriétés tels que l'indice de Jaccard [16], le coefficient de Dice [10] ou l'indice de Tversky [33]. L'avantage de cette stratégie est qu'elle évalue à la fois les points communs et les différences d'ensembles de propriétés comparées qui permettent d'exploiter plus de connaissances sémantiques que l'approche basée sur le chemin. Cependant, la limitation est qu'il est nécessaire d'équilibrer la contribution de chaque propriété en décidant la standardisation et la pondération des paramètres sur chaque propriété.

En mesurant les similitudes sémantiques basées sur le contenu de l'information (CI), on utilise le contenu de l'information comme mesure de l'information en associant des probabilités d'apparition à chaque classe ou instance dans l'ontologie et en calculant le nombre d'occurrences de ces classes ou instances dans l'ontologie [32]. De cette manière, les classes ou instances peu fréquentes deviennent plus informatives que les classes ou instances fréquentes. Un inconvénient de cette stratégie est qu'elle exige des ontologies larges avec une structure taxonomique détaillée afin de bien différencier les classes.

Au-delà de la mesure des similarités sémantiques mentionnée ci-dessus, il existe un certain nombre d'approches basées sur des combinaisons des trois principales stratégies. Par exemple, Hu et al. [14] utilisent la combinaison de la stratégie basée sur les caractéristiques et la stratégie basée sur le chemin. Ils utilisent la logique de description pour représenter les caractéristiques des entités et la mesure de similarité cosinus pour calculer une similarité. De leur côté, Batet et al. [5] utilisent l'équation 1 pour calculer la similarité sémantique basée sur les caractéristiques des classes et des instances et l'approche basée sur le contenu de l'information.

$$Sim(c_1, c_2) = -log_2 \frac{|T(c_1) \cup T(c_2)| - |T(c_1) \cap T(c_2)|}{|T(c_1) \cup T(c_2)|} \quad (1)$$

où $T(c_i) = \{c_j \in C \mid c_j$ est la superclasse de $c_i\}$, C contient la hiérarchie complète des concepts ou la taxonomie de l'ontologie.

Dans nos travaux, nous avons choisi de travailler sur la représentation d'une ontologie au moyen de triplets. Un triplet RDF comporte trois composants : sujet, prédicat et objet. En particulier, le sujet peut être le nom d'une classe, ou un instance. Le prédicat est le nom d'une propriété d'un classe ou d'un instance. L'objet est une valeur d'une propriété de la classe ou du instance qui peut se séparer en un littéral ou un nom d'une autre classe ou un autre instance. Le nom d'une classe, d'un instance, ou des littéraux sont exprimés via un texte pouvant comporter plusieurs mots. Afin de préparer leur traitement, ces contenus textuels sont vectorisés. Nous précisons dans la section suivante les méthodes que nous avons étudiées à cette fin.

2.4 Représentations vectorielles de mots

Une ontologie est composée de concepts et de relations. Ces éléments sont étiquetés par des textes (un ou plusieurs mots). Pour que les machines comprennent et effectuent des calculs sur ces contenus textuels, il faut les transformer en une représentation numérique en utilisant un corpus textuel [6]. La vectorisation de mots permet de représenter un mot par un vecteur à valeurs réelles et ce vecteur décrit le mieux possible le sens de ce mot dans son contexte. En général, plusieurs techniques sont proposées pour vectoriser un mot telles que celles basées sur la fréquence de mots (e,g. TF-IDF [31]) ou le sac de mots continus (CBOW) ou encore le saut de gramme (Skip-Gram) (e,g. Word2vec [25]).

Le TF-IDF [3] est une mesure statistique basée sur un corpus de documents [4]. Cette technique évalue la pertinence d'un

3. TF-IDF (Term Frequency-Inverse Document Frequency) est noté pour la Fréquence du Terme et la Fréquence Inverse du Document

4. Dans le contexte d'une ontologie, un ensemble de triplets est équivalent un document

mot par rapport à un document dans un corpus de documents. Tout d'abord, on calcule la fréquence relative d'un mot m dans un document d comme suit :

$$tf(m,d) = \frac{f(m,d)}{\sum_{m' \in d} f(m',d)} \tag{2}$$

où $f(m,d)$ dénote le nombre de fois où le mot m apparaît dans le document d, $\sum_{m' \in d} f(m',d)$ dénote le nombre total des mots dans le document d. Ensuite, on mesure la quantité d'informations fournies par le mot m dans le corpus de documents D avec la fréquence inverse du document comme suit :

$$idf(m,D) = log \frac{N}{|d \in D : m \in d|} + 1 \tag{3}$$

où N est le nombre de documents dans le corpus, $|d \in D : m \in d|$ est le nombre de documents où le mot m apparaît. Donc, la valeur de $tf.idf$ du mot m dans le document d au sein du corpus D est définie comme suit :

$$tf.idf(m,d,D) = tf(m,d) \times idf(m,D) \tag{4}$$

Une valeur $tf.idf(m,d,D)$ élevée d'un mot m dans un document d indique que ce mot est pertinent pour ce document au sein du corpus de documents D [31].
La technique de sac de mots continus, CBOW, construit la représentation vectorielle d'un mot m_i via la prédiction de son occurrence et la connaissance des mot avoisinants. Autrement dit, le saut de gramme, Skip-Gram, construit la représentation vectorielle d'un mot m_i en prédisant son contexte d'occurrence. Donc, étant donné une séquence de mots d'apprentissage $\{m_1, m_2, ..., m_T\}$ l'objectif du CBOW est de maximiser la moyenne des log-probabilités :

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} log\ p(m_t | m_{t+j}) \tag{5}$$

Tandis que l'objectif de Skip-gram est de maximiser la moyenne des log-probabilités :

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} log\ p(m_{t+j} | m_t) \tag{6}$$

où c est la taille du contexte. La formulation de Skip-gram définit $p(m_{t+j}|m_t)$ en utilisant la fonction softmax :

$$p(m_{t+j}|m_t) = \frac{exp((v'_{m_{t+j}})^T v_{m_t})}{\sum_{i=1}^{M} exp((v'_{m_i})^T v_{m_t})} \tag{7}$$

où v_{m_t} est la représentation vectorielle d'entrée du mot m_t, et $v'_{m_{t+j}}$, v'_{m_i} sont les représentation vectorielles de sortie du mot m_{t+j}, m_i. M est le nombre de mots dans le dictionnaire du corpus.
Word2vec est l'une des implémentations les plus populaires pour créer un plongement de mots en utilisant une architecture d'apprentissage automatique à l'aide d'un réseau de neurones. Il prédit les mots en fonction de leur contexte en combinant les deux techniques CBOW et Skip-gram [25, 4]. En particulier, la figure 1 illustre l'architecture de Word2vec qui comporte conventionnement trois couches : couche d'entrée, couche cachée, et couche de sortie. D'abord, un dictionnaire de mots avec la taille N est synthétisé à partir d'un corpus de textes. Ensuite, le processus d'apprentissage automatique crée et met à jour les valeurs des poids des matrices $W_{T \times N}$, $W'_{T \times N}$. Une fois l'apprentissage terminée, nous obtenons la matrice $W_{T \times N}$ pour le plongement de mots.

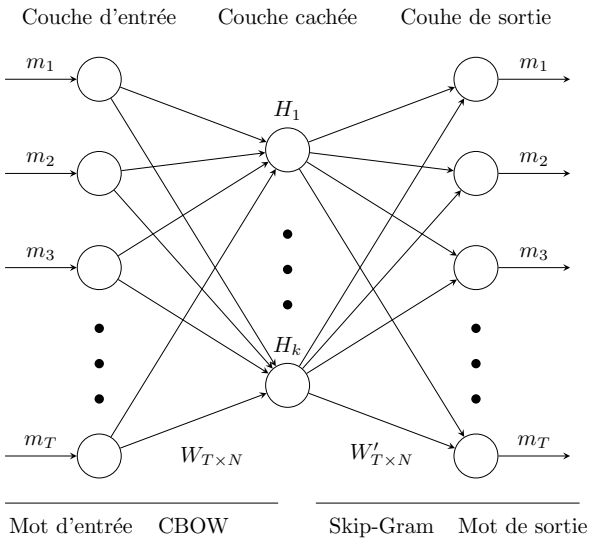


FIGURE 1 – L'architecture du modèle de Word2vec

Plusieurs plongements de mots sont créés en utilisant ce modèle pour des langues différentes [25]. Fauconnier [12], et Hadi et ses collègues [1] implémentent ce modèle à partir des textes en français. D'ailleurs, plusieurs autres travaux ont obtenu de bons résultats dans la conversion d'un mot en une représentation vectorielle tels que Fastext [7], Glove [28], le modèle Transformer avec l'implémentation du BERT [9].

Un plongement de mots entraîné avec de très grands corpus permet d'obtenir rapidement la représentation vectorielle d'un mot. Dans nos travaux, nous avons fait le choix de calculer la mesure de similarité entre deux termes textuels en tenant compte de la combinaison de CBOW et Skip-gram. La similarité entre les deux termes textuels qui se composent de mots différents peut profiter de cette forme de représentation afin de calculer la distance entre eux. Dans la section suivante, nous détaillons notre approche proposée pour mesurer de similarité au sein d'un système de recommandation.

3 Mesure de similarité au sein d'un système de recommandation

3.1 Système de recommandation pour l'achat/vente des véhicules d'occasion

Dans le cadre de nos travaux, nous nous intéressons à l'illustration de la mesure de similarité sémantique sur le système de recommandation basé sur les connaissances représentées au moyen d'ontologies dans une application e-commerce de vente/achat des véhicules d'occasion.

Les données d'un SdR basé sur la base de connaissances représentées au moyen d'ontologies se concentrent sur trois types principaux : les profils de l'utilisateur, les descriptions d'éléments ou les attributs d'éléments, et les intéractions entre les utilisateurs et les éléments. Tout d'abord, les profils d'utilisateur incluent les informations personnelles et les préférences de l'utilisateur sur les éléments de véhicule. Ils peuvent être organisés et être réécrits sous la forme des triplets formellement définis comme suit :

$$G_U = \{a_1^u, a_2^u, ..., a_n^u, \} \tag{8}$$

où a_i^u dénote le triplet $a_i^u = \langle sujet_i, prédicat_i, objet_i \rangle$. Autrement dit, le triplet a_i^u peut aussi s'exprimer comme $\langle ressource_i, propriété_i, énoncé_i \rangle$. Par exemple, "*Louis aime la voiture modèle S de Tesla*". Cette expression naturelle peut se représenter sous la forme de deux triplets différents comme $\langle Louis, aime, la_voiture_modèle_s \rangle$, $\langle la_voiture_modèle_s, est_fabiquée_par, Tesla \rangle$. Ensuite, les descriptions de véhicule peuvent également être représentées comme un graphe de connaissance. Elles peuvent être définies selon la même approche :

$$G_V = \{a_1^v, a_2^v, ..., a_n^v, \} \tag{9}$$

où a_i^v dénote le triplet $a_i^v = \langle sujet_i, prédicat_i, objet_i \rangle$ ou $a_i^v = \langle ressource_i, propriété_i, énoncé_i \rangle$. Enfin, lorsqu'un utilisateur effectue une intéraction sur des éléments de description de véhicule en donnant une note, un commentaire ou en ajoutant à une liste de favoris, on marque ces intéractions pour avoir une analyse de l'intention et du comportement de l'utilisateur afin de proposer des recommandations pertinentes. Donc, les intéractions sont définies comme une fonction à plusieurs paramètres :

$$SR : G_U \times G_V \times G_{C_1} \times ... \times G_{C_n} \rightarrow Intéraction \tag{10}$$

où G_U correspond à l'utilisateur, G_V correspond aux éléments de description de véhicule, G_{C_i}s correspond aux informations contextuelles, par exemple : objectifs, locations, temps, ressources [3]. Les ontologies sont développées pour profiler des utilisateurs et modéliser des éléments de description de véhicules [19]. Sur la base de ces ontologies, les données RDFs sont collectées et stockées dans un triplestore interrogeable au moyen de requêtes SPARQL. Des règles peuvent être définies pour déduire ou filtrer les éléments en utilisant les inférences ontologies. Dans ce cas, le SdR basé sur les connaissances comporte les quatre principales tâches suivantes :

— Recevoir et analyser les demandes des utilisateurs à partir de l'interface utilisateur.
— Construire et réaliser des requêtes sur la base de connaissance.
— Calculer des similarités sémantiques entre l'élément de véhicules, le profil utilisateur.
— Classer les éléments correspondant aux besoins de l'utilisateur.

Les mesures de similarité entre les éléments ou le profil utilisateur est une tâche importante pour générer la liste des recommandations la plus pertinente. Le travail s'effectue à partir des données RDFs qui sont organisées sous la forme de triplets $\langle sujet, prédicat, objet \rangle$. Les comparaisons entre deux triplets se limitent souvent aux objets communs ou non communs. Les informations de sujet et prédicat peuvent cependant également fournir des informations importantes sur l'objet lui-même et sa comparaison avec d'autres triplets. Dans la section suivante nous présentons comment dans notre approche nous exploitons ces deux accès à l'information pour calculer les similarités sémantiques entre les triplets d'une base de connaissances.

3.2 Mesure de similarité sémantique entre les triplets

Nous avons choisi de définir une approche hybride tenant compte de la combinaison des approches de calcul de la mesure de similarité sémantique basées sur les caractéristiques et basées sur le contenu de l'information. Le sujet, le prédicat et l'objet dans un triplet contiennent des informations importantes. Un ensemble de triplets permet d'agréger des informations provenant de triplets simples. Par conséquent, la mesure de la similarité sémantique entre ensembles de triplets doit prendre en compte tous les triplets/éléments de chaque ensemble.

La mesure de la similarité sémantique se concentre sur la comparaison de deux ensembles de triplets à partir de tous leurs éléments en les séparant en informations quantitatives et informations qualitatives. D'une part, la comparaison d'objets est réalisée en utilisant la stratégie de similarité sémantique basée sur les propriétés. D'autre part, la comparaison des sujets et des prédicats est effectuée par la stratégie de similarité sémantique basée sur le contenu de l'information.

3.3 Mesure des informations qualitatives

Les informations qualitatives font référence aux mots, aux étiquettes utilisés pour décrire les classes, les relations, et les annotations. Dans un triplet, le sujet et le prédicat expriment une information qualitative. Les objets peuvent contenir des informations qualitatives ou quantitatives. Par exemple, nous avons trois triplets suivants :
$\langle ford_focus_4_2018, la_boîte_de_vitesse, mécanique \rangle$
$\langle ford_focus_4_2020, la_boîte_de_vitesse, mécanique \rangle$
$\langle citron_c5_aircross, la_boîte_de_vitesse, mécanique \rangle$
Tous les composants de ces trois triplets sont qualitatifs. L'information du sujet de trois triplets peut être utilisée pour contribuer à la mesure de similarité entre eux. Dans cette section, nous nous concentrons sur la mesure de la simila-

rité sémantique pour les Sujets, Prédicats et Objets Qualitatifs (SPOQ). Nous proposons la même formule pour les trois composants afin de calculer la similarité.

Soient deux SPOQs a_{s1} et a_{s2} dont les vecteurs de mots sont $M_1 = \{m_{11}, m_{12}, ..., m_{1k}\}$ et $M_2 = \{m_{21}, m_{22}, ..., m_{2l}\}$, leur similarité sémantique est définie comme suit :

$$Sim_1(a_{s1}, a_{s2}) = \frac{\sum_{i=1}^{k} \bar{S}(m_{1i}, a_{s2}) + \sum_{j=1}^{l} \bar{S}(m_{2j}, a_{s1})}{k+l} \tag{11}$$

où $\bar{S}(m, a_s)$ dénote la similarité sémantique d'un mot m et d'un SPOQ. La fonction $\bar{S}(m, a_s)$ est formellement calculée comme suit :

$$\bar{S}(m, a_s) = \max_{m_i \in M} \bar{S}(m, m_i) \tag{12}$$

où $m_i \in M = \{m_1, m_2, ..., m_k\}$ est le vecteur de mots de a_s. Chaque mot m_i est représenté par un vecteur numérique. On peut utiliser les techniques introduits dans le section 2.4. L'approche basée sur la fréquence de mots TF-IDF facilite l'obtention de la probabilité d'un mot dans un ensemble de triplets. Cependant, le principal inconvénient de cette approche est qu'elle ne peut pas capturer l'information sémantique du mot et l'ordre du mot dans l'ensemble de triplets parce qu'elle crée le vecteur basé sur la fréquence du mot dans un ensemble de triplets et la collection des ensembles de triplets. Nous proposons l'utilisation de CBOW et Skip-gram avec l'implémentation de Word2vec [25, 1] afin de surmonter cela. Nous calculons finalement la similarité entre deux mots m_i, m_j par la similarité cosinus : $\bar{S}(m_i, m_j) = \frac{m_i . m_j}{\|m_i\| \|m_j\|}$.

3.4 Mesure des informations quantitatives

Les informations quantitatives sont des informations numériques qui sont utilisées pour exprimer l'information de type nominal, ordinal, intervalle, ou ratio. Dans un triplet, l'objet utilise souvent cette forme d'information pour manifester des informations des propriétés pour les classes, concepts de l'ontologie. Par exemple, nous avons des triplets suivants :
$\langle ford_focus_4_2018, a_le_kilométrage, 107351 \rangle$
$\langle ford_focus_4_2020, a_le_kilométrage, 25040 \rangle$
$\langle citron_c5_aircross, a_le_kilométrage, 48369 \rangle$
Les objets de ces triplets sont des valeurs numériques. La comparaison entre chiffres s'effectue simplement par les mesures de distance. Afin de comparer deux objets différents, nous utilisons la distance euclidienne entre deux objets. Ainsi, plus la différence entre deux objets est petite, plus la similitude entre eux est grande. Soient deux objets a_{o1} et a_{o1} dont les vecteurs sont $a_{o1} = \{o_{11}, o_{12}, ..., o_{1k}\}$ et $a_{o2} = \{o_{21}, o_{22}, ..., o_{2k}\}$, leur similarité sémantique est définie comme suit :

$$Sim_2(a_{o1}, a_{o2}) = \frac{1}{1 + \sqrt{\sum_{i=0}^{k} (o_i - o_j)^2}} \tag{13}$$

3.5 Mesure des triplets

La comparaison de deux triplets $a_1 = \langle a_{s1}, a_{p1}, a_{o1} \rangle$ et $a_2 = \langle a_{s2}, a_{p2}, a_{o2} \rangle$ est effectuée en fonction du type d'information des objets dans les triplets. Si l'objet contient des informations qualitatives, la similarité sémantique entre a_1 et a_2 est définie comme suit :

$$Sim_I(a_1, a_2) = \frac{1}{N} \sum_{i \in P, \omega \in Q} \omega \times Sim_1(a_{i1}, a_{i2}) \tag{14}$$

où $P = \{s, p, o\}$ correspond aux informations de $sujet$, $predicat$, et $objet$ sous la forme de vecteur de mots. $Q = \{\alpha, \beta, \gamma\}$ est le poids respectifs pour les composants de triplet. N est le nombre de composants de triplet.

Par ailleurs, si l'objet contient des informations quantitatives, la mesure de similarité sémantique des triplets a_1 et a_2 est définie comme suit :

$$Sim_{II}(a_1, a_2) = \frac{1}{N} (\sum_{i \in P, \omega \in Q} \omega \times Sim_1(a_{i1}, a_{i2}) + \gamma \times Sim_2(a_{o1}, a_{o2})) \tag{15}$$

où $P = \{s, p\}$ correspond aux informations de $sujet$ et $predicat$ sous la forme de vecteur de mots. $Q = \{\alpha, \beta\}$ représente les poids respectifs du sujet et du prédicat. Et γ est le poids pour l'objet.

Par conséquent, la similarité sémantique de deux ensembles de triplets $G_1 = \{a_1, a_2, ..., a_g\}$ et $G_2 = \{a_1, a_2, ..., a_g\}$ est calculée sur la base de comparaison de similarité de chaque triplet simple comme suit :

$$Sim(G_1, G_2) = \frac{1}{L} (\sum_{j=0}^{L} Sim_I(a_{1j}, a_{2j})) + \frac{1}{H} (\sum_{j=0}^{H} Sim_{II}(a_{1j}, a_{2j})) \tag{16}$$

où L est le nombre de triplets qui contient les objets qualitatifs. H est le nombre de triplets qui contient les objets quantitatifs.

4 Cas expérimental

Dans cette section nous testons notre approche dans le cas d'une application d'achat/vente de véhicules. Nous mesurons ainsi la similarité sémantique entre deux ensembles de triplets représentant chacun un véhicule. Tout d'abord, la transformation d'un mot à un vecteur est réalisée en utilisant le corpus de mots entraîné qui est développé dans le travail de Hadi et al [1]. Nous avons choisi d'utiliser le modèle CBOW et Skip-gram au lieu de TF-IDF à cause des problèmes concernant la capture de l'information sémantique qui est presque impossible sur la technique de TF-IDF. La figure 2 démontre la distance très proche des mots, groupes de mots dans un même secteur en utilisant les vecteurs entraînés de Word2vec.

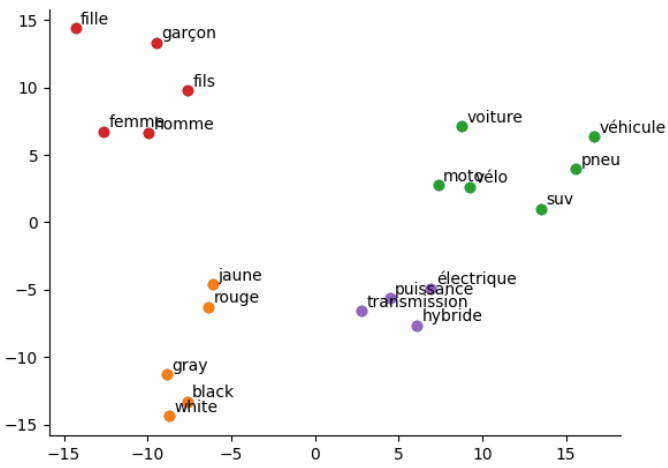


FIGURE 2 – La distance proche entre les mots, groupes qui sont vectorisés par le corpus de mots entraîné

En utilisant l'ontologie, nous pouvons reconstruire la base de connaissances d'un domaine sous une forme lisible par des machines ainsi que les humaines. À partir des ontologies des véhicules développées dans le travail [19], nous réalisons une collection des instances des classes et leurs relations afin de créer un triplestore de données RDF. La figure 3 illustre deux ensembles de triplets représentant deux voitures.

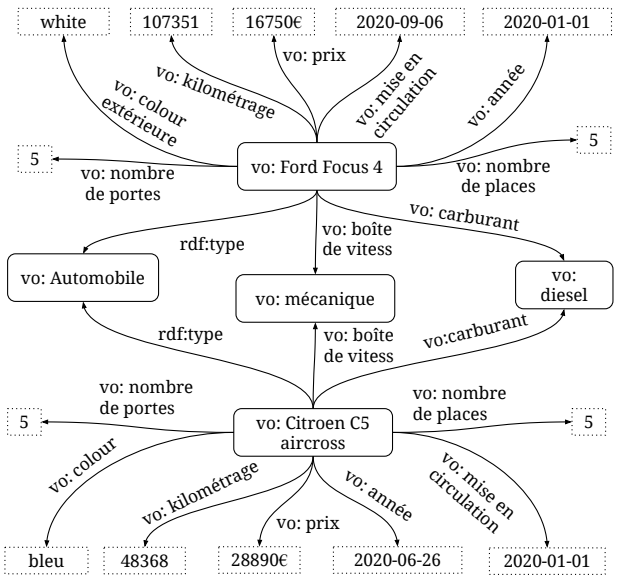


FIGURE 3 – Les données de triplets sont représentées par un graphe (*vo* note pour l'Ontologie de Véhicule)

La description de chaque véhicule est représentée par les informations textuelles et les informations numériques. Notre approche propose de séparer la mesure de la similarité en deux calculs séparés. L'un s'applique aux informations numériques parce qu'elles exigent les calculs simples pour avoir la distance ou la similarité. L'autre s'applique aux informations textuelles lorsqu'une classe où un instance d'une ontologie se composent à partir d'un groupe de mots et chaque mot a des dépendance sémantique avec les autres. En profitant des travaux du domaine de traitement du langage naturel avec les méthodes d'apprentissage profond sur de très grands corpus, les données catégorielles peuvent être représentées dans un vecteur numérique qui contient les relations du mot avec les mots récurant provenant de plusieurs documents en ligne.

		V_1	V_2	V_3	V_4	V_5
V_1	SiLi	1.0				
	N-2	1.0				
	N-1	1.0				
V_2	SiLi	0.57	1.0			
	N-2	0.50	1.0			
	N-1	**0.61**	1.0			
V_3	SiLi	0.50	0.48	1.0		
	N-2	0.48	0.46	1.0		
	N-1	**0.64**	**0.57**	1.0		
V_4	SiLi	0.54	0.49	**0.62**	1.0	
	N-2	0.49	0.47	0.50	1.0	
	N-1	**0.58**	**0.60**	0.59	1.0	
V_5	SiLi	0.54	0.46	**0.69**	0.68	1.0
	N-2	0.48	0.45	0.53	0.52	1.0
	N-1	**0.59**	**0.55**	0.60	**0.71**	1.0

TABLE 1 – La mesure de similarité entre les 5 voitures avec les trois approches différentes

Sur la base des instances collectées, nous réalisons des expérimentations et des évaluations sur trois approches suivantes :

1. N-1 : notre approche proposée principale avec l'utilisation du modèle de Word2vec [1] pour vectoriser les informations qualitatives.
2. N-2 : notre approche avec l'utilisation du modèle de TF-IDF pour vectoriser les informations qualitatives.
3. SiLi : l'approche proposée par Siying Li et ses collèges [21], cette approche hybride combine la stratégie basée sur le contenu et celle sur les caractéristiques mais ne considère que les objets et les prédicats des triplets.

La table 1 affiche les résultats de calcul de la similarité entre les 5 voitures de marques différentes. En particulier, V_1 est le "*Renault captur 2*", V_2 est la marque "*posrche taycan*", V_3 est le "*ford focus 4*", V_4 est la marque "*audi a1 sportback*", et V_5 est le "*citroen c5 aircross*". Les ensembles de triplets de ces voitures sont montrés dans l'appendice A.

En analysant les résultats obtenus et présentés dans la table 1, nous arrivons sur plusieurs conclusions. Premièrement, notre approche **N-1** donne le résultat de calcul de la similarité entre les voitures plus élevé que les autres approches dans 8 sur 10 cas de comparaisons. Toutefois, le résultat de calcul de la similarité de notre approche est moins élevé que l'approche de **SiLi** dans la comparaison de deux cas : entre les voitures V_4, V_3 et entre les voitures V_5, V_3. Deuxièmement, notre approche en utilisant le technique TF-IDF **N-2** pour la représentation vectorielle de mot a obtenu les résultats le plus bas dans tout les cas de comparaison. Cela s'explique par la capacité de capture des informations contextuelles et sémantiques de l'approche Word2vec qui est meilleur que celle de l'approche TF-IDF.
Les expérimentations montrent que notre approche **N-1** a obtenu de bons résultats pour les mesures de similarité entre les ensembles de triplets. L'utilisation du sujet dans la comparaison permet d'ajouter de l'information à la mesure de similarité d'un triplet. Aussi, la distinction contenus textuels et numériques permet d'appliquer la formule appropriée selon le type de contenu. Au final la somme des deux calculs représente la similarité mesurée. Compte tenu des cette distinction, de la prise en compte des triplets contextuels et du calcul à partir des contenus textuels enrichi des dépendances sémantiques entre les mots constituant le texte, la similarité obtenue est plus précise que celles rencontrées dans la littérature [32, 24, 21].

5 Conclusion et perspectives

La mesure de similarité sémantique sur la base de l'ontologie est une tâche importante pour proposer une liste de recommandations pertinentes à un utilisateur. Dans cet article, nous proposons une stratégie hybride qui combine la stratégie basée sur les caractéristiques et basée sur le contenu de l'information. Avec notre approche, afin de ne pas perdre d'information, les trois composants d'un triplet sont considérés dans le calcul de similarité. La distinction de type de données, textuel ou numérique, permet d'effectuer un calcul adapté et plus précis. Nous avons effectué une première expérimentation de notre approche et l'avons comparée à deux autres calculs de similarité. Les résultats obtenus montrent son intérêt. Nous devons maintenant poursuivre nos travaux et en premier lieu effectuer d'autres tests sur des corpus différents et des applications différentes. Nous devons concéder que les mots non considérés dans le corpus entraîné posent un problème. En perspective, la résolution de ce problème ainsi que la construction d'un corpus des triplets entraînés pourraient être des travaux prometteurs dans le futur.

Références

[1] Hadi Abdine, Christos Xypolopoulos, Moussa Kamal Eddine, and Michalis Vazirgiannis. Evaluation of word embeddings from large-scale french web content. 2021.

[2] Serge Abiteboul, Ioana Manolescu, Philippe Rigaux, Marie-Christine Rousset, and Pierre Senellart. *Ontologies, RDF, and OWL*, page 143–170. Cambridge University Press, 2011.

[3] Gediminas Adomavicius and Alexander Tuzhilin. *Context-Aware Recommender Systems*, pages 217–253. Springer US, Boston, MA, 2011.

[4] Oussama Ahmia, Nicolas Béchet, Pierre-François Marteau, and Alexandre Garel. Utilité d'un couplage entre word2vec et une analyse sémantique latente : expérimentation en catégorisation de données textuelles. In *Extraction et Gestion des Connaissances : Actes de la conférence EGC*, 2019.

[5] Montserrat Batet, David Sánchez, and Aida Valls. An ontology-based measure to compute semantic similarity in biomedicine. *Journal of biomedical informatics*, 44(1) :118–125, 2011.

[6] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in Neural Information Processing Systems*, 13, 2000.

[7] Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov, Matthijs Douze, and Herve Jegou. Fasttext.zip : Compressing text classification models. *arXiv preprint arXiv :1612.03651*, 2016.

[8] Richard Cyganiak, David Wood, Markus Lanthaler, Graham Klyne, Jeremy J Carroll, and Brian McBride. Rdf 1.1 concepts and abstract syntax. *W3C recommendation*, 25(02) :1–22, 2014.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*, 2018.

[10] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3) :297–302, 1945.

[11] Yu Du, Sylvie Ranwez, Nicolas Sutton-Charani, and Vincent Ranwez. Apports des ontologies aux systèmes de recommandation : état de l'art et perspectives. In *30es Journées Francophones d'Ingénierie des Connaissances, IC 2019*, pages 64–77, 2019.

[12] Jean-Philippe Fauconnier. French word embeddings, 2015.

[13] N Guarino, P Giaretta, and N Mars. Towards very large knowledge bases : Knowledge building and knowledge sharing. ontologies and knowledge bases : Towards a terminological clarification. n. *Mars. Amsterdam, IOS Press*, pages 25–32, 1995.

[14] Bo Hu, Yannis Kalfoglou, Harith Alani, David Dupplaw, Paul Lewis, and Nigel Shadbolt. Semantic metrics. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 166–181. Springer, 2006.

[15] Mohammed E Ibrahim, Yanyan Yang, David L Ndzi, Guangguang Yang, and Murtadha Al-Maliki. Ontology-based personalized course recommendation framework. *IEEE Access*, 7 :5180–5199, 2018.

[16] Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37 :547–579, 1901.

[17] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. *Knowledge-based recommendation*, page 81–123. Cambridge University Press, 2010.

[18] Rui Jiang, Mingxin Gan, and Peng He. Constructing a gene semantic similarity network for the inference of disease genes. In *BMC systems biology*, volume 5, pages 1–11. Springer, 2011.

[19] Ngoc Luyen Le, Marie-Hélène Abel, and Philippe Gouspillou. Towards an ontology-based recommender system for the vehicle sales area. In Luigi Troiano, Alfredo Vaccaro, Nishtha Kesswani, Irene Díaz Rodriguez, and Imene Brigui, editors, *Progresses in Artificial Intelligence & Robotics : Algorithms & Applications*, pages 126–136, Cham, 2022. Springer International Publishing.

[20] Wei-Nchih Lee, Nigam Shah, Karanjot Sundlass, and Mark Musen. Comparison of ontology-based semantic-similarity measures. In *AMIA annual symposium proceedings*, volume 2008, page 384. American Medical Informatics Association, 2008.

[21] Siying Li, Marie-Hélène Abel, and Elsa Negre. Ontology-based semantic similarity in generating context-aware collaborator recommendations. In *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 751–756. IEEE, 2021.

[22] Jie Lu, Dianshuang Wu, Mingsong Mao, Wei Wang, and Guangquan Zhang. Recommender system application developments : A survey. *Decision Support Systems*, 74 :12–32, 2015.

[23] LE Ngoc Luyen, Anne Tireau, Aravind Venkatesan, Pascal Neveu, and Pierre Larmande. Development of a knowledge system for big data : Case study to plant phenotyping data. In *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics*, pages 1–9, 2016.

[24] Rouzbeh Meymandpour and Joseph G Davis. A semantic similarity measure for linked data : An information content-based approach. *Knowledge-Based Systems*, 109 :276–293, 2016.

[25] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*, 2013.

[26] Van Nguyen. Ontologies and information systems : a literature survey. 2011.

[27] Charbel Obeid, Inaya Lahoud, Hicham El Khoury, and Pierre-Antoine Champin. Ontology-based recommender system in higher education. In *Companion Proceedings of the The Web Conference 2018*, pages 1031–1034, 2018.

[28] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove : Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[29] Catia Pesquita, Daniel Faria, Andre O Falcao, Phillip Lord, and Francisco M Couto. Semantic similarity in biomedical ontologies. *PLoS computational biology*, 5(7) :e1000443, 2009.

[30] M Andrea Rodriguez and Max J. Egenhofer. Determining semantic similarity among entity classes from different ontologies. *IEEE transactions on knowledge and data engineering*, 15(2) :442–456, 2003.

[31] Gerard Salton and Michael J McGill. Introduction to modern information retrieval. 1986.

[32] David Sánchez, Montserrat Batet, David Isern, and Aida Valls. Ontology-based semantic similarity : A new feature-based approach. *Expert systems with applications*, 39(9) :7718–7728, 2012.

[33] Amos Tversky. Features of similarity. *Psychological review*, 84(4) :327, 1977.

[34] Giannis Varelas, Epimenidis Voutsakis, Paraskevi Raftopoulou, Euripides GM Petrakis, and Evangelos E Milios. Semantic similarity methods in wordnet and their application to information retrieval on the web. In *Proceedings of the 7th annual ACM international workshop on Web information and data management*, pages 10–16, 2005.

A Appendice : Ensembles de triplets des voitures utilisés dans les expérimentations

```
<vo:RC2,rdf:type,vo:Automobile>
<vo:RC2,vo:année,vo:2022-01-01>
<vo:RC2,vo:mis en circulation,vo
    :2022-04-28>
<vo:RC2,vo:contrôle technique,vo:non
    requis>
<vo:RC2,vo:kilométrage,vo:5493>
<vo:RC2,vo:carburant,vo:hybride essence é
    lectrique>
<vo:RC2,vo:boîte de vistesse,vo:
    automatique>
<vo:RC2,vo:coleur extérieure,vo:noir>
<vo:RC2,vo:nombre de portes,vo:5>
<vo:RC2,vo:nombre de places,vo:5>
<vo:RC2,vo:puissance fiscale,vo:5>
<vo:RC2,vo:puissance din,vo:93>
<vo:RC2,vo:Critique d'Air,vo:1>
<vo:RC2,vo:emission de CO2,vo:35>
<vo:RC2,vo:consommation mixte,vo:1.5>
<vo:RC2,vo:norme euro,vo:euro6>
<vo:RC2,vo:fabriquer par,vo:Renault
    occasion>
<vo:RC2,vo:type de véhicule,vo:4x4, SUV &
    Crossover occasion>
<vo:RC2,vo:location,vo:Cher>
```

```
<vo:RC2,vo:price,vo:36580>
```

Listing 1 – V_1 Renault Captur 2 (RC2)

```
<vo:PT,rdf:type,vo:Automobile>
<vo:PT,vo:année,vo:2022-01-01>
<vo:PT,vo:mis en circulation,vo
    :2022-09-10>
<vo:PT,vo:contrôle technique,vo:non requis
    >
<vo:PT,vo:kilométrage,vo:4932>
<vo:PT,vo:carburant,vo:electrique>
<vo:PT,vo:boîte de vistesse,vo:automatique
    >
<vo:PT,vo:coleur intérieure,vo:cuir ivoire
    >
<vo:PT,vo:coleur extérieure,vo:noir metal>
<vo:PT,vo:nombre de portes,vo:4>
<vo:PT,vo:nombre de places,vo:4>
<vo:PT,vo:garranty,vo:20 mois>
<vo:PT,uvso:puissance fiscale,vo:8>
<vo:PT,vo:puissance din,vo:530>
<vo:PT,vo:Critique d'Air,vo:0>
<vo:PT,vo:emission de CO2,vo:0>
<vo:PT,vo:norme euro,vo:euro6>
<vo:PT,vo:fabriquer par,vo:Porsche
    occasion>
<vo:PT,vo:type de véhicule,vo:Berline
    occasion>
<vo:PT,vo:location,vo:Rhône>
```

Listing 2 – V_2 Porsche Taycan (PT)

```
<vo:FF4,rdf:type,vo:Automobile>
<vo:FF4,vo:année,vo:2020-01-01>
<vo:FF4,vo:mis en circulation,vo
    :2020-09-06>
<vo:FF4,vo:contrôle technique,vo:non
    requis>
<vo:FF4,vo:kilométrage,vo:107351>
<vo:FF4,vo:carburant,vo:diesel>
<vo:FF4,vo:boîte de vistesse,vo:mécanique>
<vo:FF4,vo:coleur extérieure,vo:gris foncé
    >
<vo:FF4,vo:nombre de portes,vo:5>
<vo:FF4,vo:nombre de places,vo:5>
<vo:FF4,vo:garranty,vo:12 mois>
<vo:FF4,vo:puissance fiscale,vo:4>
<vo:FF4,vo:puissance din,vo:95>
<vo:FF4,vo:Critique d'Air,vo:2>
<vo:FF4,vo:emission de CO2,vo:89>
<vo:FF4,vo:consommation mixte,vo:4.5>
<vo:FF4,vo:norme euro,vo:euro6>
<vo:FF4,vo:fabriquer par,vo:Ford occasion>
<vo:FF4,vo:type de véhicule,vo:Berline
    occasion>
<vo:FF4,vo:location,vo:Loiret>
<vo:FF4,vo:price,vo:16750>
```

Listing 3 – V_3 Ford Focus 4 (FF4)

```
<vo:AA1,rdf:type,vo:Automobile>
<vo:AA1,vo:année,vo:2018-01-01>
<vo:AA1,vo:mis en circulation,vo
    :2018-09-15>
<vo:AA1,vo:contrôle technique,vo:non
    requis>
<vo:AA1,vo:kilométrage,vo:20211>
<vo:AA1,vo:carburant,vo:diesel>
<vo:AA1,vo:boîte de vistesse,vo:
    automatique>
<vo:AA1,vo:coleur extérieure,vo:bleu>
<vo:AA1,vo:coleur intérieure,vo:noir>
<vo:AA1,vo:nombre de portes,vo:5>
<vo:AA1,vo:nombre de places,vo:5>
<vo:AA1,vo:garranty,vo:12 mois>
<vo:AA1,vo:puissance fiscale,vo:4>
<vo:AA1,vo:puissance din,vo:90>
<vo:AA1,vo:Critique d'Air,vo:2>
<vo:AA1,vo:emission de CO2,vo:101>
<vo:AA1,vo:consommation mixte,vo:3.6>
<vo:AA1,vo:norme euro,vo:euro6>
<vo:AA1,vo:fabriquer par,vo:Audi occasion>
<vo:AA1,vo:type de véhicule,vo:Citadine
    occasion>
<vo:AA1,vo:location,vo:Yvelines>
<vo:AA1,vo:price,vo:23200>
```

Listing 4 – V_4 Audi A1 sportback (AA1)

```
<vo:CC5,rdf:type,vo:Automobile>
<vo:CC5,vo:année,vo:2020-01-01>
<vo:CC5,vo:mis en circulation,vo
    :2020-06-26>
<vo:CC5,vo:contrôle technique,vo:non
    requis>
<vo:CC5,vo:kilométrage,vo:48368>
<vo:CC5,vo:carburant,vo:diesel>
<vo:CC5,vo:boîte de vistesse,vo:mécanique>
<vo:CC5,vo:coleur extérieure,vo:bleu>
<vo:CC5,vo:nombre de portes,vo:5>
<vo:CC5,vo:nombre de places,vo:5>
<vo:CC5,vo:garranty,vo:12 mois>
<vo:CC5,vo:puissance fiscale,vo:6>
<vo:CC5,vo:puissance din,vo:131>
<vo:CC5,vo:Critique d'Air,vo:2>
<vo:CC5,vo:emission de CO2,vo:106>
<vo:CC5,vo:consommation mixte,vo:4.1>
<vo:CC5,vo:norme euro,vo:euro6>
<vo:CC5,vo:fabriquer par,vo:Citroen
    occasion>
<vo:CC5,vo:type de véhicule,vo:4x4, SUV &
    Crossover occasion>
<vo:CC5,vo:location,vo:Yvelines>
<vo:CC5,vo:price,vo:28890>
```

Listing 5 – V_5 Citroen C5 aircross (CC5)

Comparaison des solutions de NLU sur un corpus français pour un chatbot de support COVID-19

Marion Schaeffer[1] et Christophe Bouvard[1]

[1] Wikit, Lyon, France

marion@wikit.ai et christophe@wikit.ai

Résumé

Les chatbots sont de plus en plus déployés au sein des organisations afin de répondre à des requêtes utilisateur·rice·s en temps réel. Ils utilisent un moteur de compréhension du langage (Natural Language Understanding) en charge d'assimiler les phrases utilisateur·rice·s et ainsi transformer l'information implicite en information explicite interprétable par la machine. De nombreuses offres de ce type existent sur le marché, et il peut être compliqué de choisir une technologie plutôt qu'une autre.

Nous proposons une comparaison de moteurs de NLU disponibles pour la langue française, suivant des critères de performance et de confiance dans un objectif de faciliter l'industrialisation. Nous utilisons deux jeux de données en français sur le cas d'usage du support à destination des employé·e·s : plus de 1000 phrases annotées concernant la COVID-19 que nous rendons disponibles en libre accès et un ensemble de phrases annotées concernant les ressources humaines.

Mots-clés

Comparaison, agent conversationnel, chatbot, traitement automatique du langage naturel (TALN), compréhension du langage naturel, classification d'intentions, reconnaissance d'entités, français, support aux employé·e·s, COVID-19.

Abstract

Chatbots have been increasingly deployed into organizations in order to provide real-time answers to user requests. They leverage a Natural Language Understanding engine to understand sentences and transform implicit information into explicit information for the machine. There are many similar offerings on the market, and choosing one technology over another can be complicated.

We propose a comparison of available NLU engines for the French language, based on criteria like performance and confidence with the aim of facilitating industrialization. We use two French datasets on the employee support use case : more than 1000 annotated sentences about COVID-19 freely available and a set of annotated sentences about human resources.

Keywords

Benchmarking, chatbot, natural language processing (NLP), natural language understanding (NLU), intent classification, entity detection, French, employee support, COVID-19.

1 Introduction

Au cours des dix dernières années, les chatbots sont devenus de plus en plus présents sur les sites web que nous consultons, ainsi que sur les applications que nous utilisons [17]. Ces programmes informatiques ont pour but d'interagir avec un·e utilisateur·rice lors d'une conversation qui se veut naturelle, c'est-à-dire telle que celle que l'on pourrait avoir avec un humain. Au cours d'un échange textuel avec des phrases exprimées en langage naturel, le chatbot doit apporter une réponse à une demande de l'utilisateur·rice [6]. L'assistance et le support utilisateur sont donc des applications privilégiées des chatbots au sein de diverses organisations telles que des entreprises ou des administrations publiques [8].

Pour comprendre la demande de l'utilisateur·rice, les chatbots utilisent un moteur de compréhension du langage naturel (NLU). L'algorithme de compréhension est un composant central du chatbot qui classifie des intentions et reconnaît des entités. Les intentions sont des catégories de sujets que le chatbot peut traiter [21] et les entités sont des mots-clés identifiés pour un traitement spécifique [10]. Les informations structurées issues du moteur NLU sont alors exploitables par le gestionnaire de dialogue du chatbot pour animer l'interaction avec l'humain. De nombreuses solutions de NLU sont disponibles sur le marché, par exemple : Dialogflow de Google (https://cloud.google.com/dialogflow), Wit de Facebook (https://wit.ai), LUIS de Microsoft (https://www.luis.ai), Amazon Lex d'AWS (https://aws.amazon.com/fr/lex/), Watson Assistant d'IBM (https://www.ibm.com/fr-fr/products/watson-assistant), SiriKit d'Apple (https://developer.apple.com/documentation/sirikit), Rasa NLU de Rasa (https://rasa.com/), Snips NLU de Snips (https://github.com/snipsco/snips-nlu), etc. Certains de ces outils sont largement configurables pour fonctionner avec des paramètres personnalisés. D'autres outils se veulent simples d'utilisation et proposent des plateformes clés en main non modifiables [4].

Lors du développement d'un chatbot, le choix de la technologie pour assurer la compréhension du langage n'est pas

une tâche aisée. Il existe différentes études comparatives et les résultats varient en fonction du *dataset*, c'est-à-dire du domaine d'application et des données en elles-mêmes (type de phrases, vocabulaire, périmètre, ...) [3]. De plus, la majorité des comparaisons sont faites en anglais [1, 5]. Pour travailler dans d'autres langues comme le français, les ressources évaluant ce type de solutions sont rares et difficiles à trouver. C'est pour cela que nous présentons une analyse comparative des moteurs de compréhension sur un jeu de données en français. Aussi, nous souhaitons évaluer les performances de ces solutions sur différents critères : la pertinence des résultats et les scores de confiance associés qui sont des critères essentiels pour l'industrialisation du chatbot. Nous avons choisi de réaliser cette comparaison sur un cas d'application concret issu d'une expérience client : un chatbot de support pour les employé·e·s pendant la pandémie de COVID-19. En effet, durant la pandémie, une grande partie de la population s'est retrouvée isolée, engendrant ainsi un sentiment de confusion et de questionnement chez de nombreux·euses salarié·e·s. L'objectif de ce jeu de données est donc de construire un chatbot pour informer les employé·e·s d'une organisation sur les dispositions spécifiques liées à la COVID-19 [2]. Pour compléter l'analyse, nous avons également effectué la comparaison sur un autre corpus traitant un cas d'usage assez proche : le support pour les employé·e·s sur les questions courantes concernant les ressources humaines.

L'article est structuré comme suit : la Section 2 présente l'état de l'art sur les moteurs de compréhension du langage et l'utilisation des chatbots durant la pandémie de COVID-19. Ensuite, la Section 3 aborde les étapes nécessaires à la réalisation de notre étude comparative. Les jeux de données utilisés sur le cas d'usage du support aux employé·e·s sur la thématique des ressources humaines et lors de la pandémie de COVID-19 sont décrits, ainsi que les solutions de NLU choisies pour l'étude. Puis, la Section 4 présente les métriques utilisées et les résultats obtenus par les moteurs de compréhension du langage suivant différents critères. Enfin, la Section 5 conclut notre étude en résumant notre contribution et en citant des perspectives d'évolution pour de futurs travaux.

2 État de l'art

2.1 Comparaison de moteurs de compréhension du langage

Des publications récentes ont comparé les services de NLU dans différents domaines et suivant diverses motivations. Dans [4], les outils des principaux fournisseurs industriels de plateformes NLU sont comparés. Les critères de comparaison sont variés : facilité d'utilisation, langues supportées, entités et intentions pré-construites, intention par défaut, intégration en ligne ou encore le coût financier d'utilisation.

De multiples jeux de données sont utilisés afin d'évaluer les technologies de NLU. En effet, les résultats varient d'un domaine d'application à l'autre. Dans [3], LUIS de Microsoft est le plus performant sur tous les jeux de données. Juste derrière se trouve Rasa, puis IBM Watson et enfin Dialogflow. Un autre article ajoute la technologie Snips au benchmark [5]. Elle se classe à la deuxième place, juste derrière LUIS et devant Rasa avec le même ensemble de données. D'autre part, la technologie d'IBM est la plus performante sur des corpus différents : [1, 20]. Les jeux de données varient de par leur taille (nombre d'intentions et nombre de phrases d'exemples par intention), mais aussi de par leur domaine d'application (vocabulaire et formulations de phrase). Les discussions s'attardent rarement sur l'explication des variations de performances d'un corpus à l'autre, ou sur la justification des scores de confiance.

Le marché étant en constante évolution, la liste des solutions de NLU change d'une étude comparative à l'autre. Nous avons choisi de présenter six des principaux acteurs du marché permettant de travailler en français, ainsi qu'une méthode qui est un point de comparaison (*baseline*) pour la tâche de classification : les Machines à Vecteurs de Support (SVM).

2.1.1 IBM Watson

Watson Assistant est l'agent virtuel intelligent d'IBM. Il s'agit d'une plateforme conversationnelle qui permet la compréhension du langage (NLU) mais aussi la gestion du dialogue (*Dialog Manager*). L'outil supporte plusieurs langues et permet une gestion en ligne ou via une API. Aussi, la partie NLU comprend les intentions et interprète les entités en évaluant la question de l'utilisateur·rice à partir de la base de connaissances disponible. Un score de confiance est ainsi attribué aux prédictions faites.

2.1.2 Rasa

Rasa propose un service de NLU open source. Il permet aux développeur·euse·s de configurer, de déployer et d'exécuter le moteur NLU sur des serveurs en local ou déployés en production. De multiples configurations sont disponibles, avec des paramètres par défaut mais aussi la possibilité d'intégrer des outils tels que spaCy ou BERT par exemple. La solution gère les intentions, les entités et les dialogues avec des scores associés aux résultats. Elle offre une adaptabilité, un contrôle des données et donc des avantages importants pour une solution déployée directement au sein de l'entreprise.

Les configurations incluant des modèles de langue pré-entraînés de type BERT sont particulièrement intéressantes pour bénéficier de la fonctionnalité multilingue [7, 19].

2.1.3 Classification par SVM avec la représentation du langage pré-entraînée CamemBERT

Les machines à vecteurs de support sont des modèles d'apprentissage supervisé utilisés pour des problèmes de discrimination ou de régression. Comme présenté dans [21], les SVM cherchent à définir une frontière de décision entre deux classes en utilisant une fonction noyau. Cette frontière doit avoir une marge maximale dans un espace latent.

Pour classifier les requêtes utilisateur·rice·s, il faut d'abord transformer le texte en vecteur. La méthode de représentation du langage pré-entraînée CamemBERT peut être utilisée à cet escient. CamemBERT [16] est un modèle de langage français basé sur l'architecture RoBERTa [14] et la

partie française du corpus OSCAR [18]. Il permet d'obtenir des vecteurs représentatifs et contextualisés au sens sémantique.

2.1.4 Snips

L'écosystème Snips permet de construire des assistants vocaux à partir d'une console web. Un moteur de compréhension du langage parlé (*Spoken Language Understanding*) est composé d'un moteur de reconnaissance automatique de la parole (*Automatic Speech Recognition*) et d'un moteur de compréhension du langage naturel (NLU). Le modèle peut être entraîné en anglais, français et allemand [5]. Pour la partie classification d'intentions et reconnaissance d'entités, des scores de confiance sont également attribués aux prédictions.

2.1.5 Wit

La plateforme Cloud de NLU wit.ai est détenue et maintenue par Facebook. L'utilisation de la plateforme est gratuite et plusieurs langues y sont disponibles. La solution se concentre uniquement sur l'extraction du sens de l'énoncé d'un·e utilisateur·rice avec la classification d'intentions et la détection d'entités et ne gère aucun type de conversation ou d'intégration via des plateformes de messagerie. L'outil fournit des scores de confiance.

2.1.6 LUIS

Le service Language Understanding, communément appelé LUIS, est l'une des briques de l'offre Microsoft pour construire des systèmes avec Intelligence Artificielle conversationnelle. LUIS utilise des algorithmes d'apprentissage pour analyser les requêtes des utilisateur·rice·s. Les intentions et les entités sont prédites avec des scores de confiance. Plusieurs langues sont disponibles, ainsi qu'un accès via son portail personnalisé, des API et des bibliothèques de développement (dont SDK appelé Bot Framework) pour compléter LUIS lors de l'implémentation de chatbot.

2.1.7 Dialogflow

La plateforme de création d'agents conversationnels Dialogflow de Google permet d'analyser des entrées textuelles ou audio pour en extraire du sens. La compréhension du langage naturel de Dialogflow utilise des modèles fondés sur BERT. De nombreuses langues sont disponibles pour analyser les intentions et les entités avec des scores de confiance ainsi que pour la gestion du dialogue.

2.2 Les langages moins représentés que l'anglais

La plupart des résultats sur les comparaisons de solutions de NLU sont obtenus à partir de corpus écrits en anglais. L'article [22] est un travail pionnier qui aborde ce problème pour la langue italienne. À notre connaissance, il est très difficile d'obtenir des informations similaires pour la langue française. Les difficultés sont multiples : seuls les prestataires disposant d'outils français peuvent être comparés (c'est actuellement le cas pour une majorité). De plus, il est nécessaire de créer un jeu de données spécifique à la tâche et à la langue car il n'existe pas de *datasets* français standardisés comme c'est le cas en anglais, par exemple avec le corpus de dialogue Ubuntu [15, 3].

2.3 L'utilisation des chatbots durant la pandémie de COVID-19

Depuis le début de la pandémie, les recherches sur la COVID-19 se multiplient, notamment en informatique. Les chatbots sont un moyen de centraliser les informations et de les rendre disponibles en continu. C'est pourquoi ils ont été largement utilisés pendant la crise sanitaire, dans différentes langues, sur divers canaux de communication et pour différentes applications [9]. À titre d'illustration, nous pouvons mentionner :

— le suivi des patient·e·s [12], pour collecter simplement des informations sur l'évolution des symptômes prolongés,
— le dépistage des employé·e·s du système de santé [11], pour éviter la propagation nosocomiale de l'infection,
— la recherche d'information [13, 2], pour fournir des réponses vérifiées aux demandes des utilisateur·rice·s et ainsi lutter contre la désinformation.

Les moteurs de compréhension du langage ont donc été largement étudiés sur divers cas d'application mais principalement en langue anglaise. De plus, les chatbots ont montré un réel intérêt durant la crise sanitaire engendrée par la pandémie de COVID-19. La suite de l'article détaille donc notre apport pour la comparaison de solutions de compréhension du langage, en français, sur le cas d'usage de la COVID-19 au sein des entreprises.

3 Méthodologie

3.1 Corpus COVID-19

La comparaison des services de compréhension du langage est basée sur le corpus d'un chatbot en production dédié aux questions concernant la COVID-19 au sein d'une collectivité territoriale française. Les utilisateur·rice·s finaux·ales de cet agent conversationnel sont les employé·e·s de cette organisation, qui ont accès à des services numériques dans le cadre de leur travail. Cet outil d'aide à l'information a été mis en œuvre pour répondre aux thématiques rencontrées lors des premiers confinements.

L'objectif de ce chatbot est de soutenir les employé·e·s lorsqu'ils·elles rencontrent des problèmes liés à la crise sanitaire et de les aider dans leurs interrogations quotidiennes concernant la COVID-19 (tests PCR, gel désinfectant pour les mains, masques, vaccins, cas contact, travail à distance, etc.).

Le jeu de données d'entraînement et de test en français sur le cas d'usage de la COVID-19 utilisé pour évaluer les performances des moteurs de compréhension du langage est disponible à l'adresse suivante : https://github.com/wikit-ai/nlu-french-benchmark.

Les phrases d'entraînement ont été majoritairement exportées du chatbot en production. De ce fait, elles ont été rédigées par les clients et correspondent donc exactement à

leurs besoins. Le corpus a ensuite été complété par des experts afin d'égaliser le nombre de phrases par intention et de garantir les meilleures performances possibles pour chacune des solutions de NLU. Les experts en question travaillent dans le domaine du chabot depuis plusieurs années et leur quotidien consiste à optimiser l'utilisation du chatbot du point de vue fonctionnel et de la compréhension du langage. Au total, 330 phrases ont été annotées pour 22 intentions différentes.

Le corpus de test a été conçu uniquement par les experts de façon manuelle. Pour chaque intention, ces derniers ont rédigé trois phrases différentes par intention en intégrant des entités afin que les différentes classes et mots-clés soient testés dans les mêmes proportions. L'expérience des annotateur·rice·s leur permet de s'approcher au mieux de la façon dont les utilisateur·rice·s discutent et surtout du type de requêtes qui sont généralement utilisées. Les doublons au sein du corpus de test et entre le corpus d'entraînement et de test ont ensuite été retirés. Au final, 913 phrases sont annotées.

3.1.1 Le cas d'usage de la COVID-19

La crise sanitaire de la COVID-19 a bouleversé la vie quotidienne des travailleur·euse·s. Le télétravail a été présenté comme une réorganisation de la vie professionnelle, avec de nouveaux outils et de nouvelles règles. Des processus spécifiques ont été mis en place avec des changements rapides et fréquents en réponse à la propagation du virus. Ainsi, ils·elles ont beaucoup de questions à poser au quotidien [2, 13].

Par exemple, voici quelques énoncés d'utilisateur·rice·s issus du jeu de données d'entraînement du chatbot en production précédemment présenté :

- *J'ai été en contact avec une personne qui a été testée positive au COVID-19.*
- *Quels sont les gestes barrière ?*
- *Je suis très angoissée par cette pandémie.*
- *Peut-on travailler à distance ?*
- *Je dois acheter du gel désinfectant pour les mains.*
- *Quels types de masques peuvent être utilisés au bureau ?*
- *Quelles sont les règles en vigueur à la cantine de l'entreprise ?*

Les réponses à ces entrées utilisateur·trice sont des données informatives pour guider les employé·e·s dans les processus et instructions qui évoluent rapidement.

3.1.2 Les intentions

La classification d'intentions consiste à interpréter la requête de l'utilisateur·rice en fonction des connaissances du système. En effet, il existe une liste exhaustive d'intentions que le moteur NLU est capable de reconnaître. Par conséquent, le problème est de savoir quelle intention existante est la plus proche sémantiquement de la phrase de l'utilisateur·rice [21].

Nous avons conservé 22 intentions différentes qui couvrent un large éventail de questions que les employé·e·s d'une organisation peuvent se poser pendant la pandémie de COVID-19.

Les intentions sont listées dans le tableau 1. Chacune a été entraînée avec 15 phrases d'exemples variées.

Index	Intentions du corpus COVID-19
1	Comment les absences sont-elles gérées ?
2	Comment se passent les repas au restaurant d'entreprise ?
3	Découvrir le champ d'action du bot
4	Que faire en situation de cas contact ou avéré dans l'équipe ?
5	Explique-moi les gestes barrières
6	Quelles sont les dispositions pour que je puisse garder mes enfants ?
7	J'ai une question sur la cellule psychologique
8	J'ai une question sur la prise de repas en salle commune
9	J'ai une question sur le travail à distance
10	J'ai une question sur les campagnes de dépistage
11	J'ai une question sur les campagnes de vaccination
12	J'ai une question sur les cas contacts
13	J'ai une question sur les formations
14	J'ai une question sur les masques
15	J'ai une question sur les symptômes
16	J'ai été testé·e positif·ve à la COVID-19
17	Les réunions en présentiel sont-elles autorisées ?
18	Obtenir du gel hydroalcoolique
19	Obtenir une attestation de déplacement pendant le couvre-feu
20	Est-ce qu'on peut prendre le café entre collègues ?
21	Quel est le dispositif pour les agents vulnérables ?
22	Retrouver tous les modes opératoires

TABLE 1 – Intentions du jeu de données sur la COVID-19

3.1.3 Les entités

La reconnaissance d'entités nommées (*NER*) est également une tâche très importante pour les chatbots. Elle consiste à identifier les portions de texte qui désignent des entités nommées telles que des personnes, des lieux, des noms d'organisations, ... Pour notre cas d'usage, nous étendons cette détection à des entités au sens large afin de pouvoir reconnaître des noms de logiciels ou des mots-clés liés au cas d'usage.

Notre comparaison intègre donc les méthodes d'extraction d'entités sur des entités personnalisées listées dans le tableau 2 afin d'améliorer la classification des intentions. Ces entités peuvent être spécifiées de manière exhaustive avec une liste de synonymes associés ou de termes connexes. En effet, les auteurs de [1] ont démontré que les intentions contenant des mots exclusifs et des entités distinctes étaient plus faciles à identifier par tous les moteurs NLU. Cela peut s'expliquer par le fait que les systèmes de NLU utilisent les types d'entités extraits comme entrée pour la classification

des intentions, mais aussi et surtout parce que certains mots sont associés à des intentions spécifiques.

Entités	Synonymes et termes connexes
absence	congé, RTT, vacances
agent	employé, collègue
attestation	attestation de déplacement, attestation dérogatoire
café	chocolat, chocolat chaud, choco, thé, eau chaude
cellule psychologique	cellule psy, soutien psychologique, soutien psy
compte épargne temps	CET
conjoint	conjointe, mari, femme, époux, épouse
dépistage	PCR, test, test antigénique
enfant	bébé, nourrisson, fils, fille
établissement	collège, crèche, école, école élémentaire, école maternelle, maternelle, lycée
gel	gel hydro, gel hydroalcoolique, hydroalcoolique, spray désinfectant, désinfectant
gestes barrières	règles, règles sanitaires
HDD	département, Hôtel du département
masque	FFP2, masque chirurgical, masque FFP2
mode opératoire	dispositif, dispositif sanitaire, mesure sanitaire, mode opé, protocole, protocole sanitaire
psychologue	psy
rendez-vous	rdv, rendez vous
restaurant administratif	restaurant, resto, resto admin, resto administratif
réunion	réunions, réu, meeting
trad	télétravail, télé travail, télé-travail, travail à distance, travail à domicile
véhicule	voiture

TABLE 2 – Entités du jeu de données sur la covid-19

3.2 Corpus RH

Afin de conforter les tendances dégagées grâce au corpus traitant le cas d'usage de la COVID-19, nous comparons les performances des différentes solutions de NLU sur un autre corpus. Ce corpus regroupe des intentions et des entités relatives aux ressources humaines dans le domaine de l'entreprise.

Le dataset a été conçu de façon incrémentale par les experts de la société Wikit et leurs clients afin de correspondre au mieux à leurs besoins. Seules 15 intentions ont été sélectionnées pour ce test, et chaque intention est entraînée avec 12 phrases d'exemple. La partie test rassemble 8 phrases différentes par intention.

Le corpus couvrent les congés, les périodes d'essai, les formations en entreprise, le comité social et économique de l'entreprise, la mutuelle, les interlocuteurs des ressources humaines, les questions relatives aux salaires, les risques de burnout ou encore le télétravail. Les interrogations autour de ces sujets sont converties en intentions, et les mots-clés sont utilisés comme entités.

3.3 Les plateformes de compréhension du langage

Les algorithmes de NLU visent à extraire des informations utiles et structurées à partir de données non structurées, c'est-à-dire d'une entrée en langage naturel. Nous sélectionnons six services largement utilisés par les chercheurs et les entreprises proposant une analyse en français afin de comparer leurs performances. Ces NLU apparaissent également dans d'autres études à des fins de comparaison mais avec des domaines d'application différents [1, 4, 5, 22, 20].

3.3.1 Interrogation via API

Pour les plateformes disponibles dans le Cloud comme Watson Assistant, LUIS, Dialogflow et Wit, les modèles de compréhension du langage sont entraînés et interrogés grâce à des API. L'entraînement est réalisé en associant chaque phrase d'exemple à l'intention correspondante et en annotant chaque mot-clé avec son entité. Suivant les plateformes, l'entraînement est automatique (Watson Assistant) ou doit être déclenché (Dialogflow, LUIS, Wit). Finalement, la prédiction sur le jeu de test se fait phrase par phrase avec le classement des intentions associées à leurs scores ainsi que les entités trouvées et leur position.

3.3.2 Création de modèles en local

En ce qui concerne Rasa, Snips et la classification par SVM, des modèles sont créés en local et sont testés directement en local également. Tout comme pour les plateformes disponibles dans le Cloud, les phrases d'exemples sont associées à leur intention et il en va de même pour les entités. Le modèle est entraîné (non automatique) puis requêté afin de prédire les intentions et entités d'une nouvelle phrase. Les intentions sont classées et associées à un score. Les entités sont identifiées avec leur position dans la phrase.

4 Résultats

Dans cette section, nous présentons la comparaison des NLU en termes de classification d'intentions et de détection d'entités. Nous entraînons chacun des services de NLU avec les corpus en français, puis nous testons la solution avec les phrases de test annotées.

4.1 Classification d'intentions

Pour évaluer la performance des moteurs de compréhension du langage sur la tâche de classification d'intentions, nous ne considérons que l'intention classée en première position, c'est-à-dire l'intention ayant le score de confiance le plus élevé. En effet, lors d'une conversation réelle avec le chatbot, celui-ci ne fournit qu'une seule réponse en fonction de la meilleure intention trouvée. C'est donc cette intention candidate qui doit être évaluée.

4.1.1 Métriques

Différentes métriques peuvent être utilisées pour évaluer les performances des NLU. Elles fournissent des informations complémentaires sur les performances du système. Afin de comprendre ces métriques, nous définissons pour une classe A fixée :

- les *vrais positifs (TP)* comme étant les valeurs appartenant à la classe A et effectivement prédites comme telles
- les *vrais négatifs (TN)* comme étant les valeurs n'appartenant pas à la classe A et effectivement prédites comme telles
- les *faux positifs (FP)* comme étant les valeurs qui n'appartiennent pas à la classe A mais qui sont prédites comme y appartenant
- les *faux négatifs (FN)* comme étant les valeurs appartenant à la classe A mais qui ne sont pas prédites comme y appartenant.

Exactitude : L'exactitude (ou l'*accuracy*) est la métrique la plus largement utilisée. Elle divise le nombre d'observations correctement prédites par le nombre total d'observations, comme le montre l'équation 1. Cependant, cette mesure peut être insuffisante si l'ensemble de données n'est pas symétrique.

$$Exactitude = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

Précision : Dans le domaine de la recherche d'information, la précision est la fraction des documents retrouvés qui sont pertinents pour la requête. Il s'agit d'une évaluation quantitative de la performance. Ainsi, comme nous pouvons le voir dans l'équation 2, la précision est le rapport entre les observations d'une classe A correctement prédites et le total des observations prédites comme étant de classe A.

$$Précision = \frac{TP}{TP + FP} \tag{2}$$

Rappel : Toujours dans le domaine de la recherche d'information, le rappel (ou *recall*) est la fraction des documents pertinents qui sont retrouvés avec succès. Il s'agit d'une évaluation qualitative de la performance. Ainsi, comme nous pouvons le voir dans l'équation 3, le rappel est le rapport entre les observations d'une classe A correctement prédites et l'ensemble réel des observations de la classe A.

$$Rappel = \frac{TP}{TP + FN} \tag{3}$$

Score F1 : Le score F1 est la moyenne harmonique de la précision et du rappel. Cette métrique reflète au mieux la qualité d'un modèle, en cas de distribution inégale des classes par exemple, car elle prend en compte les faux positifs et les faux négatifs comme le présente l'équation 4.

$$\begin{aligned} \textit{Score F1} &= 2 \times \frac{Rappel \times Précision}{Rappel + Précision} \\ &= \frac{2 \times TP}{2 \times TP + FP + FN} \end{aligned} \tag{4}$$

4.1.2 Classement

Nous comparons les moteurs de compréhension du langage avec les scores F1 obtenus sur le jeu de test COVID-19 et RH. Comme présenté dans le tableau 3 pour le corpus COVID-19, Wit a le meilleur score F1 d'une valeur de $0,806$. La performance de Watson est proche avec un score de $0,794$. Juste après, le modèle de classification par SVM associé à la méthode d'*embedding* CamemBERT et Rasa obtiennent un score très proches de $0,782$ et $0,780$. LUIS et Snips se placent respectivement en cinquième et sixième position avec des scores de $0,756$ et $0,746$. Enfin, Dialogflow se positionne à la fin du classement avec un score de $0,668$.

NLU	Exactitude	Précision	Rappel	Score F1
Watson	**0,800**	0,812	**0,800**	0,794
Rasa	0,783	0,794	0,783	0,780
SVM	0,784	0,795	0,785	0,782
Snips	0,731	0,815	0,729	0,746
Wit	0,752	**0,891**	0,753	**0,806**
LUIS	0,766	0,785	0,766	0,756
Dialogflow	0,617	0,812	0,616	0,668

TABLE 3 – Performance de la classification d'intentions sur le corpus COVID-19

En ce qui concerne le corpus RH, les résultats sont présentés dans le tableau 4. Watson se place largement en tête avec un score de $0,923$. Juste derrière, Rasa obtient un score de $0,902$. Wit, LUIS et Snips se placent respectivement en troisième, quatrième et cinquième position avec des scores de $0,891$, $0,880$ et $0,876$. Dialogflow et le SVM se partage le bas du classement avec des scores de $0,844$ et $0,808$.

NLU	Exactitude	Précision	Rappel	Score F1
Watson	0,9	**0,959**	0,9	**0,923**
Rasa	**0,908**	0,912	**0,908**	0,902
SVM	0,817	0,841	0,817	0,808
Snips	0,875	0,899	0,875	0,876
Wit	0,85	0,961	0,85	0,891
LUIS	0,883	0,899	0,883	0,880
Dialogflow	0,808	0,929	0,808	0,844

TABLE 4 – Performance de la classification d'intentions sur le corpus RH

4.1.3 Score de confiance

Les résultats de la classification d'intentions étant assez similaires d'une technologie à une autre, il est intéressant de comparer les scores de confiance attribués aux prédictions pour chacun des modèles et des corpus.

En effet, un modèle efficace doit avoir des scores de confiance élevés sur des prédictions correctes, avec un écart-type petit, alors que le score de confiance doit être le plus bas possible pour les mauvaises prédictions, avec un écart-type plus élevé.

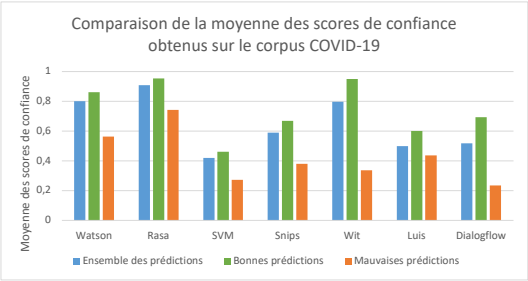


FIGURE 1 – Moyenne des scores de confiance obtenus sur le corpus COVID-19

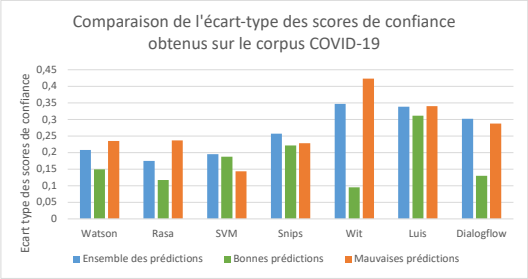


FIGURE 2 – Écart-type des scores de confiance obtenus sur le corpus COVID-19

Comme on peut le voir sur les figures 1 et 2, le comportement attendu est bien présent sur les services Watson Assistant, Rasa, Wit et Dialogflow. Il est toutefois notable que sur Snips, Dialogflow, LUIS et de façon accentuée sur le SVM, les scores obtenus sont plus bas que les scores des autres services. De même, les écart-types des bonnes et des mauvaises prédictions du SVM, de Snips et de LUIS sont très proches en comparaison avec les autres solutions.
Pour le corpus sur le cas d'usage des ressources humaines, les résultats sont présentés sur les figures 3 et 4. Les observations sont très similaires à celles faites sur le corpus COVID-19. On remarque que les scores de confiance de Rasa sont très élevés, tout comme pour Wit. Watson Assistant et Dialogflow ont également des scores assez élevés. Les scores de confiance les plus faibles sont attribués par le SVM, Snips et LUIS.
Concernant les écart-types, le comportement attendu (écart-type plus élevé pour les mauvaises prédictions que pour les prédictions correctes) est bien présent sur Rasa, Wit, LUIS et Dialogflow. Pour Watson, les écart-types des prédictions correctes et mauvaises sont très proches. Finalement, les écart-types des bonnes prédictions sont supérieurs aux écart-types des mauvaises prédictions pour le SVM et Snips.

4.2 Détections d'entités

Dans un second temps, nous comparons les moteurs de compréhension du langage avec les performances obtenues sur la détection d'entités. Pour cela, l'exactitude est mesurée, c'est à dire la proportion d'entités correctement identifiées parmi l'ensemble des entités présentes. Seules les entités présentes dans les phrases du corpus de test sont évaluées.

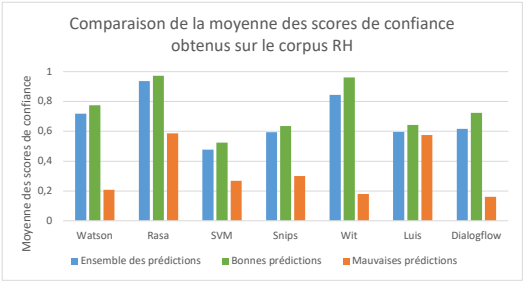


FIGURE 3 – Moyenne des scores de confiance obtenus sur le corpus RH

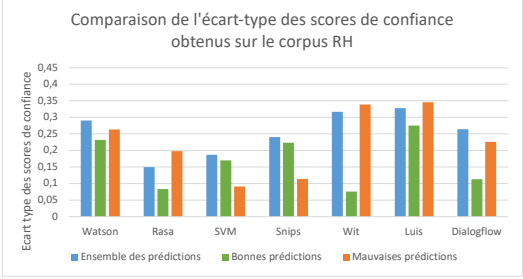


FIGURE 4 – Écart-type des scores de confiance obtenus sur le corpus RH

Comme présenté sur la figure 5 pour le corpus COVID-19, Dialogflow a les meilleures performances avec un score de $0,876$. Watson est juste derrière avec un score de $0,849$. La suite du classement est occupée par Wit avec un score de $0,747$. LUIS, Rasa et finalement Snips se positionnent au bas du classement avec des scores inférieurs.
Le SVM n'a pas été testé sur la tâche de reconnaissance d'entités. Pour implémenter notre propre algorithme de détection d'entités, il est possible, par exemple, d'utiliser les champs aléatoires conditionnels (*conditional random fields*, CRF).
En ce qui concerne le corpus RH, les résultats sont présentés en figure 6. La meilleure performance est réalisée par Dialogflow avec un score de $0,957$. LUIS se positionne juste derrière avec un score de $0,932$. Les autres solutions se suivent avec des scores assez proches, dans l'ordre il y a Rasa, Watson, Snips et un peu plus loin Wit.

5 Discussion et conclusion

Dans la section 4, nous examinons les performances des différents services de NLU sur un corpus traitant le cas d'usage de la COVID-19. Nous croisons également nos résultats avec des tests réalisés grâce au dataset du cas d'usage des ressources humaines.

5.1 Classification d'intentions et détection d'entités

Globalement les performances des services de NLU sont meilleures sur le corpus RH que sur le corpus COVID-19

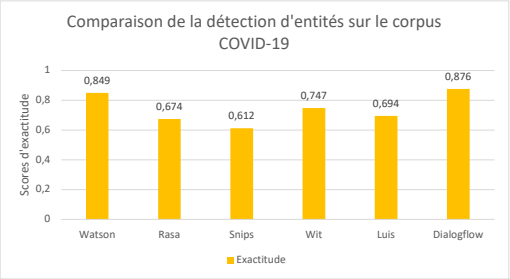


FIGURE 5 – Résultats de la détection d'entités pour le corpus COVID-19

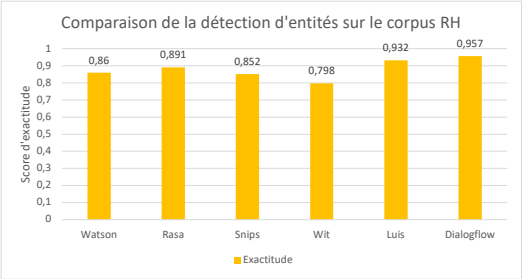


FIGURE 6 – Résultats de la détection d'entités pour le corpus RH

simplement car la quantité de données de test du corpus COVID-19 (plus de 900 phrases) est supérieure à celle du corpus RH (120 phrases).
En ce qui concerne la classification d'intentions, les moteurs de compréhension ont des scores F1 très proches. Watson semble cependant prendre l'avantage, suivi de très près part Wit et Rasa. Sur les deux corpus, Dialogflow, LUIS et Snips se positionnent en fin de classement pour cette tâche. Lors de cette comparaison, nous utilisons l'annotation d'entités lorsque les plateformes disposent de la fonctionnalité directement (c'est-à-dire toutes les plateformes sauf la méthode de classification par SVM). D'autres études telles que [1] ont montré que la présence d'entités bénéficie à la classification des intentions. Finalement, le SVM reste concurrent aux autres méthodes de classification malgré son désavantage.
En ce qui concerne la détection d'entités, Dialogflow se positionne nettement devant ses compétiteurs.
De plus, Rasa et le classifieur SVM ont un avantage supplémentaire car il est possible de les configurer et de mieux ajuster leurs paramètres pour améliorer les résultats. Par exemple, il serait intéressant d'affiner un modèle de représentation du langage pour mieux s'adapter au vocabulaire spécifique du cas d'usage du support aux employé·e·s durant la pandémie de COVID-19 ou sur de domaine des ressources humaines ; et ainsi mesurer le bénéfice de ces méthodes par rapport à d'autres services plus statiques.

5.2 Synthèse des caractéristiques

Le tableau 5 rassemblent quelques unes des caractéristiques intéressantes à comparer en vue de l'industrialisation d'une solution de compréhension du langage.

5.2.1 Fonctionnalités

Différentes fonctionnalités des solutions ont été comparées dans la section 2. Nous les reprenons ici afin de conclure notre étude.

API. Le fait de disposer d'une API est un atout majeur pour une solution : en effet, elle permet d'accélérer l'industrialisation du moteur de compréhension du langage. Toutes les solutions en proposent une, plus ou moins complète, qui permet à minima de créer un modèle et de l'interroger lors de la prédiction. Les API poussées permettent également de gérer l'entraînement et la mise à jour des modèles.
Le SVM ne dispose pas d'une API initialement car celle ci doit être implémentée. C'est un inconvénient : cela prend du temps, mais c'est aussi un atout : les spécifications peuvent être personnalisées.

Multilingue. Le multilingue est une fonctionnalité très intéressante, elle permet de gérer plusieurs langues dans un seul moteur de compréhension du langage. Les seules techniques qui proposent simplement cette fonctionnalité sont Rasa et le SVM, grâce à des modèles de représentation du langage multilingue comme Bert par exemple.
Pour les autres solutions, elles proposent différentes langues mais ces langues ne peuvent pas être gérées simultanément. En d'autres termes, il est nécessaire d'avoir un moteur de compréhension du langage par langue souhaitée.

Configurabilité du NLU. En ce qui concerne la configurabilité, nous nous intéressons ici à la possibilité d'ajuster les composants du moteur de compréhension du langage. C'est le cas pour Rasa, qui permet de choisir la représentation du langage souhaitée ainsi que les technique de traitement de la langue à utiliser. Il en va de même pour le SVM, avec des paramètres ajustables et une méthode de représentation du langage au choix. Il est également possible d'ajuster les configurations de Snips concernant la partie NLU.
Les autres solutions sont très peu configurables et proposent, au mieux, des options comme le *fuzzy matching*, l'utilisation de méthodes d'apprentissage pour la détection d'entités, etc... Ce critère est finalement très lié à celui de l'open source discuté par la suite.

Hébergement proposé. L'hébergement est une partie importante de l'industrialisation, car elle impacte la mise en production du modèle de compréhension du langage. La majorité des solutions proposent l'hébergement du modèle conjointement au logiciel en tant que service (SaaS) : Watson Assistant, Wit, LUIS, Dialogflow.
En ce qui concerne le SVM, Rasa et Snips, l'hébergement est géré par l'équipe technique en charge du projet de chatbot.

Open source. Un logiciel ou une librairie est dit open source si le code source est libre d'accès, réutilisable et modifiable. Dans cette étude, c'est bien le cas pour Rasa et Snips, dont le code source est téléchargeable et adaptable. Il en est de même pour le SVM dont différentes implémentations sont disponibles ou peut être implémenté directement. Watson Assistant, Wit, LUIS et Dialogflow communiquent

Caractéristiques	Watson	Rasa	SVM	Snips	Wit	LUIS	Dialogflow
API	✓	✓	×	✓	✓	✓	✓
Multilingue	×	✓	✓	×	×	×	×
Configurabilité du NLU	×	✓	✓	✓	×	×	×
Hébergement proposé (SaaS)	✓	×	×	×	✓	✓	✓
Open source	×	✓	✓	✓	×	×	×
Tarification proposée	✓	×	×	×	×	✓	✓
Entraînement automatique	✓	×	×	×	×	×	✓

TABLE 5 – Synthèse des caractéristiques des moteurs de compréhension du langage

très peu sur l'implémentation. Seules les grandes lignes des technologies utilisées sont abordées mais il est compliqué d'obtenir ce genre de renseignements en tant qu'utilisateur·trice.

Tarification proposée. L'aspect financier de la mise en production d'un service de NLU pour le déploiement d'un chatbot est un critère de sélection important pour des entreprises, ou bien même les universitaires lorsqu'ils souhaitent utiliser ce type de technologies. Différentes tarifications existent :

— Rasa (version open source), Snips et un modèle de classification SVM sont des technologies gratuites à utiliser. Cependant, il est nécessaire de déployer une instance de production et donc de payer un hébergement. Le tarif dépend alors de l'hébergeur mais aussi des ressources nécessaires pour faire fonctionner l'instance, ainsi que de l'utilisation faite.
— LUIS, Dialogflow et Watson Assistant proposent le moteur de compréhension du langage hébergé dans des versions d'essais gratuites et avec différents plans de tarification selon le besoin. Le coût dépend alors du nombre de requêtes émises ou des options souscrites.
— Wit.ai est totalement gratuit, et ce même pour une utilisation commerciale. Cela comprend l'entraînement et l'hébergement du modèle.

Pour continuer cette étude dans de futurs travaux, nous souhaitons définir un cadre représentatif de l'utilisation de ce type de chatbot et ainsi comparer le coût financier de son déploiement sur une période et un volume de requête donné.

Entraînement automatique. Le temps d'entraînement (ou d'apprentissage) des moteurs de compréhension du langage est un argument important dans le choix de la technologie utilisée pour développer son chatbot. Par exemple, IBM Watson Assistant et Dialogflow proposent un entraînement qui se fait automatiquement et très rapidement alors que d'autres services comme LUIS ou encore Rasa, nécessitent un déclenchement de l'entraînement. Lorsque l'entraînement est terminé, un remplacement du modèle courant par un nouveau modèle est obligatoire sur Rasa, Snips ou le SVM, la prédiction est donc indisponible durant le temps de chargement du nouveau modèle. Ces contraintes sont intéressantes à étudier car leur impact est important pour l'industrialisation d'une solution de ce type et seront donc ajoutées à l'étude dans de futurs travaux.

De la même façon, le temps de prédiction des intentions et des entités est important pour l'industrialisation du chatbot. En effet, un des principaux atouts de ce dernier est le fait de converser en temps réel. Lors de notre utilisation, toutes les solutions ont permis cela. Cependant, étudier la robustesse des plateformes lorsque de nombreuses demandes sont formulées pour des questions d'industrialisation serait une démarche intéressante et complémentaire à notre étude.

5.3 Conclusion

Les chatbots sont de plus en plus populaires et, par conséquent, les services de NLU sont largement utilisés. Ces derniers constituent une pièce centrale du chatbot, car ils permettent d'interpréter les demandes utilisateur et donc de fournir les réponses les mieux adaptées. Le choix de la technologie pour un NLU est une tâche complexe. Les services ont des caractéristiques différentes et de nombreuses études comparent leurs performances. Cependant les résultats sont souvent différents selon le jeu de données et il y a très peu de ressources disponibles pour comparer les moteurs en langue française.

Avec la pandémie, les chatbots ont été largement utilisés pour répondre aux questions des utilisateur·rice·s sur les changements induits par la situation sanitaire. Dans ce contexte, nous avons créé, annoté et mis à disposition un jeu de données d'entraînement et de test sur le cas d'usage de la COVID-19 en entreprise. Nous avons entraîné et testé différents services de traitement du langage sur la tâche de classification d'intentions et de détection d'entités afin d'identifier l'outil le plus efficace pour le français.

Nous avons constaté que Watson, Wit et Rasa obtiennent les meilleures performances sur la classification d'intentions alors que Dialogflow est le plus performant pour la détection d'entités sur nos jeux de données en français. Cependant, les différents services ont des résultats assez similaires lorsqu'ils sont comparés dans leur globalité. Il est donc intéressant de s'intéresser aux avantages de chacun des moteurs en fonction de l'usage qu'il est prévu d'en faire. Dans un travail futur, nous comparerons les services de NLU sur des aspects complémentaires tels que le temps d'entraînement, le coût financier ou encore la robustesse lors du déploiement, afin d'avoir un classement objectif basé sur autant de critères que possible, pour faciliter l'industrialisation des chatbots.

Références

[1] Ahmad Abdellatif, Khaled Badran, Diego Costa, and Emad Shihab. A comparison of natural language understanding platforms for chatbots in software engineering. *IEEE Transactions on Software Engineering*, PP :1–1, 05 2021.

[2] Eslam Amer, Ahmed Hazem, Omar Farouk, Albert Louca, Youssef Mohamed, and Michel Ashraf. A proposed chatbot framework for covid-19. pages 263–268, 05 2021.

[3] Daniel Braun, Adrian Hernandez Mendez, Florian Matthes, and Manfred Langen. Evaluating natural language understanding services for conversational question answering systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 174–185, Saarbrücken, Germany, August 2017. Association for Computational Linguistics.

[4] Massimo Canonico and Luigi De Russis. A comparison and critique of natural language understanding tools. 2018.

[5] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. Snips voice platform : an embedded spoken language understanding system for private-by-design voice interfaces. *ArXiv*, abs/1805.10190, 2018.

[6] Menal Dahiya. A tool of conversation : Chatbot. *International Journal of Computer Sciences and Engineering*, 5 :158–161, 05 2017.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT : pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[8] Dario Fiore, Matthias Baldauf, and Christian Thiel. "forgot your password again ?" : acceptance and user experience of a chatbot for in-company it support. *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia*, 2019.

[9] Sviatlana Höhn and Kerstin Bongard-Blanchy. Heuristic evaluation of covid-19 chatbots. pages 131–144, 02 2021.

[10] Anran Jiao. An intelligent chatbot system based on entity extraction using rasa nlu and neural network. *Journal of Physics : Conference Series*, 1487 :012014, 03 2020.

[11] Timothy Judson, Anobel Odisho, Jerry Young, Olivia Bigazzi, David Steuer, Ralph Gonzales, and Aaron Neinstein. Case report : Implementation of a digital chatbot to screen health system employees during the covid-19 pandemic. *Journal of the American Medical Informatics Association : JAMIA*, 27, 06 2020.

[12] Hannah Lei, Weiqi Lu, Alan Ji, Emmett Bertram, Paul Gao, Xiaoqian Jiang, and Arko Barman. COVID-19 smart chatbot prototype for patient monitoring. *CoRR*, abs/2103.06816, 2021.

[13] Yunyao Li, Tyrone Grandison, Patricia Silveyra, Ali Douraghy, Xinyu Guan, Thomas Kieselbach, Chengkai Li, and Haiqi Zhang. Jennifer for covid-19 : An nlp-powered chatbot built for the people and by the people to combat misinformation. In *NLPCOVID19*, 2020.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta : A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[15] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The ubuntu dialogue corpus : A large dataset for research in unstructured multi-turn dialogue systems. *CoRR*, abs/1506.08909, 2015.

[16] Louis Martin, Benjamin Müller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. Camembert : a tasty french language model. *CoRR*, abs/1911.03894, 2019.

[17] Quim Motger, Xavier Franch, and Jordi Marco. Conversational agents in software engineering : Survey, taxonomy and challenges. *CoRR*, abs/2106.10901, 2021.

[18] Pedro Ortiz Suarez, Benoît Sagot, and Laurent Romary. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. 07 2019.

[19] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert ? *CoRR*, abs/1906.01502, 2019.

[20] Haode Qi, Lin Pan, Atin Sood, Abhishek Shah, Ladislav Kunc, and Saloni Potdar. Benchmarking intent detection for task-oriented dialog systems. *CoRR*, abs/2012.03929, 2020.

[21] Jetze Schuurmans and Flavius Frasincar. Intent classification for dialogue utterances. *IEEE Intelligent Systems*, PP :1–1, 11 2019.

[22] Matteo Zubani, Luca Sigalini, Ivan Serina, and Alfonso Gerevini. Evaluating different natural language understanding services in a real business case for the italian language. In *KES*, 2020.

Construction d'un graphe de connaissances à partir des annotations d'articles scientifiques et de leur contenu en sciences de la vie

N. Yacoubi Ayadi[1], C. Faron[1], F. Michel[1], R. Bossy[2], A. Barbe[1]

[1] University Côte d'Azur, Inria, CNRS, I3S (UMR 7271), France
[2] MaIAGE, INRAE, Université Paris-Saclay, 78350 Jouy-en-Josas, France

yacoubi@i3s.unice.fr, faron@i3s.unice.fr, fmichel@i3s.unice.fr, robert.bossy@inrae.fr, arnaud.barbe@etu.univ-cotedazur.fr

Résumé

Dans ce papier, nous présentons un graphe de connaissances RDF permettant de décrire, structurer et intégrer des annotations d'entités nommées extraites automatiquement par l'outil Alvis NLP à partir de publications scientifiques portant sur la génétique et le phénotypage de blé. Ces entités nommées se réfèrent à la fois à des noms de gènes, traits, phénotypes, marqueurs et variétés impliqués dans la culture du blé. Cependant, une fois extraites, ces annotations sont stockées dans un format brut rendant difficile leur exploitation par les chercheurs. D'où, notre intérêt de les transformer (lifter) en un format compatible avec les standards de publication de données liées afin de construire un graphe de connaissances dans lequel des entités provenant à la fois de bases de connaissances génomiques et d'articles scientifiques ont été sémantiquement décrites et intégrées. Basé sur un ensemble de questions de compétence formulées par un expert du domaine, nous avons validé la pertinence du modèle proposé et par conséquent le graphe de connaissances généré.

Mots-clés

Ontologie, Données liées, Annotation sémantique, Graphe de connaissances RDF, Fouille de textes.

Abstract

In this paper, we present an RDF knowledge graph to describe, structure and integrate annotations of named entities automatically extracted by the Alvis NLP tool from scientific publications on wheat genetics and phenotyping. These named entities refer to the names of genes, traits, phenotypes, markers and varieties involved in wheat breeding. However, once extracted, these annotations are stored in a raw format making it difficult for researchers to exploit them. Hence, our interest in transforming (lifting) them into a format compatible with linked data publication standards in order to build a knowledge graph in which knowledge coming from both genomic knowledge bases and scientific articles has been semantically described and integrated. Based on a set of competency questions formulated by a domain expert, we validated the relevance of the proposed model and consequently the generated knowledge graph.

Keywords

Ontologies, Linked Data, Semantic annotation, Knowledge Graph, Text Mining.

1 Introduction

La culture du blé est l'une des plus importantes et répandues dans le monde, elle fournit le principal apport de protéines pour une grande part de la population mondiale. Les semanciers et sélectionneurs cherchent à obtenir des variétés aux propriétés intéressantes pour la productivité, la résistance aux maladies et l'adaptation aux changements climatiques. Les techniques modernes de phénotypage et de dépistage génomique permettent une sélection ciblée et une meilleure hybridation des variétés de blé. Ces techniques rendent possible l'obtention de graines résistantes aux maladies et à la sécheresse, tout en étant productives et durables. Ces techniques combinent une recherche génétique fondamentale en laboratoire et une expérimentation sur terrain. Une partie des résultats de ces recherches est enregistrée dans des bases de données génomiques libres d'accès. En revanche, une autre partie n'est accessible qu'à travers l'exploration de la littérature scientifique. Cependant, il est impossible pour un chercheur de parcourir l'ensemble des publications scientifiques vu leur volume exponentiel (plus de 4000 sont publiées par an). Par conséquent, les techniques de TAL (Traitement Automatique de Langues naturelles) ont été largement utilisées pour la fouille de textes dans l'objectif d'extraire et de synthétiser les informations pertinentes pouvant aider les chercheurs dans leurs investigations.

Dans ce contexte, la plate-forme AlvisNLP [2] offre différents outils TAL pour l'extraction d'entités nommées à partir des publications scientifiques permettant ainsi d'annoter différents types d'entités à savoir des gènes, des phénotypes, des traits, des variétés, des marqueurs génétiques et des taxons. Toutefois, les résultats générés par cet outil sont exportés dans un format brut (i.e., CSV) ; ce qui entrave leur exploitation et intégration avec d'autres sources de connaissances.

Nous présentons, dans ce papier, le travail de recherche que

nous avons mené dans le cadre du projet D2KAB [1], un projet de recherche ANR ayant pour objectif de créer un cadre pour transformer les données d'agronomie et de biodiversité en connaissances décrites sémantiquement, interopérables, exploitables et ouvertes. L'objectif de notre travail s'aligne complètement avec l'objectif du projet D2KAB et vise la construction d'un graphe de connaissances RDF intégrant des entités provenant de différents sources et processus, à la fois des bases de connaissances génomiques et des workflows TAL appliqués à la littérature scientifique. Ainsi, ce graphe de connaissances permettra d'intégrer les annotations extraites à partir des articles scientifiques avec des d'autres ressources terminologiques et sémantiques publiées dans le cadre du Web du Web sémantique. Guidés par un ensemble de questions de compétences (CQ), nous avons proposé un modèle qui réutilise des ontologies et des vocabulaires existants pour structurer et représenter de façon uniforme à la fois les publications scientifiques et leurs méta-données et les annotations d'entités nommées dans le même graphe de connaissances. Le processus de construction du graphe est réalisé en deux phases parallèles : (1) l'utilisation de l'outil morph-xR2RML [10] pour lifter les annotations produites par l'outil AlvisNLP en RDF, (2) l'utilisation d'un micro-service SPARQL pour récupérer les méta-données descriptives des publications scientifiques à partir de l'API PMC Entrez [2] et de les intégrer avec les annotations liftées. Enfin, toutes les CQ ont été traduites en requêtes SPARQL et évaluées pour valider le graphe de connaissances obtenu et le modèle sémantique sous-jacent ; les résultats des requêtes ont été validées par les experts.

Ce papier est organisé comme suit. Dans la section 2, nous présentons une synthèse (non exhaustive) des approches de construction de graphes de connaissances à partir de textes scientifiques ; ainsi que les vocabulaires réutilisés dans ce travail de recherche. Dans la section 3, nous présentons un ensemble de CQ qui résument les exigences et les besoins potentiels des experts d'exploiter les annotations générées. Le modèle sémantique du graphe de connaissances est présenté dans la section 4. Dans la section 5, nous détaillons le processus et les outils que nous avons utilisés pour la génération du graphe de connaissances. Dans la section 6, nous présentons des requêtes SPARQL qui correspondent à l'implémentation de certains CQ présentées dans la section 3.

2 État de l'art

Dans cette section, nous discutons quelques approches existantes pour la construction de graphes de connaissances à partir de textes. Ensuite, nous présentons les vocabulaires et les ontologies que nous avons réutilisés pour structurer les connaissances dans le futur graphe, à savoir : les ontologies FaBio (the FRBR-aligned Bibliographic Ontology) [13], BIBO (BIBliographic Ontology) et le vocabulaire Web Annotation Vocabulary (OA) [14].

2.1 Construction de graphes de connaissances à partir de textes

La problématique de construction de graphes de connaissances à partir de textes a suscité l'intérêt de plusieurs communautés incluant celles du Web sémantique, des données liées et du TAL. En effet, les techniques développées dans chacun de ces domaines s'avèrent complémentaires [9]. Une approche de construction de graphes à partir de textes doit combiner plusieurs outils et techniques pour permettre : (1) l'extraction et la reconnaissance d'entités nommées [5, 12], (2) le liage des entités nommées (normalisation d'entités nommées) à des concepts existants dans des ontologies/vocabulaires du domaine, (3) l'extraction de relations [12], et (4) la génération automatique du graphe RDF (RDFisation) [1] et sa publication conformément aux des principes des données liées. Le principal défi réside dans le liage des entités nommées. En effet, dans le domaine biomédical, les vocabulaires sont souvent volumineux et complexes. De plus, on observe un décalage important entre les étiquettes de concepts et les mentions dans le texte avec notamment l'usage extensif d'abréviations, de métonymies et de variations syntaxiques ([8], [3]). Le travail de recherche présenté dans [6] décrivant le challenge BioNLP-ST 2013 a mis en évidence l'intérêt de construire des graphes de connaissances RDF. Pour ce challenge, 10 bases de connaissances RDF ont été construites et évaluées à partir des annotations extraites par 10 systèmes TAL. L'objectif était de proposer un nouvel axe pour l'évaluation de la pertinence des annotations générées par ces systèmes. Ainsi, plusieurs requêtes SPARQL ont été conçues à l'aide d'experts du domaine et évaluées en comparant leurs résultats avec ceux obtenus de la base de connaissances de référence construite partir du *gold standard*. Contrairement aux auteurs dans [6] qui ont adopté un vocabulaire minimal conçu pour le besoin du challenge, nous nous basons sur l'utilisation conjointe de vocabulaires standards (i.e, [14]) et d'ontologies (i.e., [11]) pour la modélisation RDF des annotations générées a partir des publications scientifiques.

2.2 Vocabulaires et Ontologies existantes

Pour représenter à la fois les publications scientifiques et les annotations extraites à partir de ces publications, nous avons réutilisé plusieurs vocabulaires et ontologies. D'une part, nous avons adopté les ontologies FaBio (the FRBR-aligned Bibliographic Ontology) [13] et BIBO [3] pour représenter les méta-données descriptives et bibliographiques des publications scientifiques. L'ontologie FaBio est une ontologie dérivée du modèle FRBR [4] qui est un vocabulaire RDF publié par l'IFLA (International Federation of Library Association) pour représenter les notices artistiques et bibliographiques. FRBR définit un ensemble exhaustif de classes permettant de modéliser tout type d'oeuvres et de décrire tout le cycle de vie de l'oeuvre de sa création jusqu'à son adaptation et transformation. D'autre part, L'ontologie BIBO est une ontologie minimale qui décrit es-

1. http://www.d2kab.org/
2. https://www.ncbi.nlm.nih.gov/pmc/tools/developers/
3. https://github.com/structureddynamics/Bibliographic-Ontology-BIBO

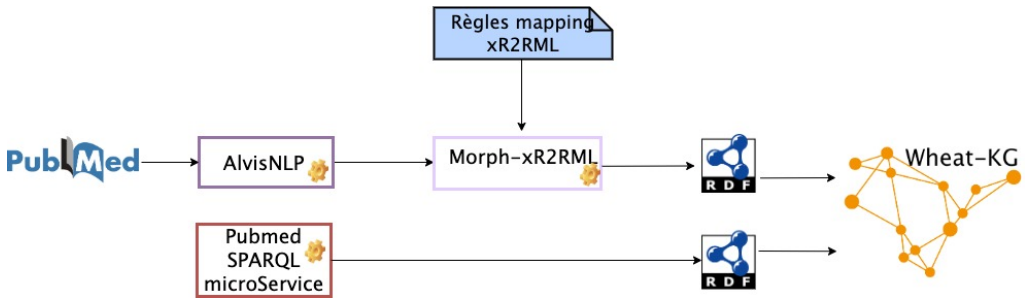


FIGURE 1 – Schéma général du pipeline de l'approche de construction du graphe Wheat-KG

sentiellement les agents, les documents et les événements qui conduisent à la production d'une œuvre. Les ontologies BIBO et FaBio sont généralement utilisées de façon complémentaire avec d'autres vocabulaires existants tels que Dublin Core [4]ou schema.org [5].

Enfin, le vocabulaire Web Annotation Vocabulary (OA) [14] est une recommandation W3C qui propose un ensemble de classes et de propriétés RDF pour représenter de manière uniforme les annotations sur le Web dans un format interopérable, d'où l'intérêt de son utilisation [7].

3 Questions de Compétences

Les questions de compétences permettent de résumer les exigences et les attentes des experts vis-a-vis d'un futur modèle de connaissances. Ces questions expriment les besoins des experts à explorer la littérature scientifique de la génomique de blé en exploitant les relations existantes entre les entités reconnues et annotées par AlvisNLP. Dans le contexte de ce travail de recherche, les CQ ont permis d'éliciter les attentes des chercheurs travaillant sur la génomique de blé et désirant exploiter la littérature scientifique autour de ce sujet. Ainsi, les CQ que nous présentons s'articulent autour des besoins des experts à explorer des possibles interactions entre les entités nommées en exploitant leur contexte de co-occurence dans le texte.

CQ1. Quelles publications scientifiques du corpus mentionnent le gène '*Lr34*' ?

CQ2. Quels sont les gènes mentionnés à proximité du phénotype *'drought tolerance'* ? Cette requête permet aux scientifiques de rechercher des gènes impliqués dans le contrôle d'un phénotype en particulier. Un ensemble de publications mentionnant un ou plusieurs gènes apparaissant avec le phénotype en question sont retournées comme résultat.

CQ3. Quels sont les marqueurs génétiques mentionnés à proximité d'un gène, qui lui-même est mentionné à proximité d'un phénotype particulier ? Cette requête permet de rechercher des publications mentionnant des marqueurs qui pourraient servir à sélectionner un phénotype donné. Comme les techniques de marquage génétique ont évolué au fil du temps et certaines sont devenues obsolètes, l'expert peut également raffiner sa requête pour sélectionner uniquement les publications apparues après 2010. D'où l'intérêt de représenter dans le graphe des méta-données descriptives telles que la date d'apparition de la publication.

CQ4. Quelles sont les variétés de blé qui présentent un phénotype particulier ? L'expert peut rechercher des variétés d'intérêt car elles présentent un phénotype spécifique.

CQ5. Effectuer une recherche bibliographique de toutes les publications mentionnant des gènes spécifiques à des variétés de blé tendre (*Triticum aestivum*) qui présentent un phénotype général. Le résultat à cette requête devra inclure une liste d'articles mentionnant à la fois des gènes, une ou plusieurs variétés de blé tendre et le phénotype en question. Cependant, si l'expert est intéressé par les gènes impliqués dans la résistance aux pathogènes, il faudrait alors inclure dans les résultats de la requête tout type de pathogènes (bactérie, virus, champignons). D'où l'intérêt d'intégrer dans le graphe des connaissances ontologiques et terminologiques qui proviennent des ontologies et des vocabulaires du domaine.

4 Modèle proposé

La définition d'un modèle qui capture la nature des entités et leurs relations dans le graphe de connaissances est impérative. En effet, le futur graphe de connaissances intégrera diffèrent types d'entités dont la sémantique sera décrite différemment selon la nature de l'entité.

4.1 Description des articles scientifiques

Pour représenter et décrire les publications scientifiques, nous avons réutilisé les vocabulaires suivants : Dublin Core, FRBR aligned Bibliographic Ontology (FaBiO) et Bibliographic Ontology (BIBO). Ces vocabulaires définissent une liste exhaustive de méta-données descriptives pour décrire les publications scientifiques telles que le DOI, l'année de publication, le nombre de pages, le journal, etc. Ainsi, un article scientifique sera représenté comme une instance des classes `fabio:ResearchPaper` et `bibo:AcademicArticle`. Les propriétés Dublin Core `dct:title` et `dct:abstract` permettent de relier la publication à son titre et son résumé. Certains résumés d'articles scientifiques sont structurés en 3 sous-sections que nous représentons comme étant 3 entités différentes identifiées chacune par un URI unique. La propriété `frbr:partOf` sera utilisée pour représenter le

4. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
5. https://schema.org/

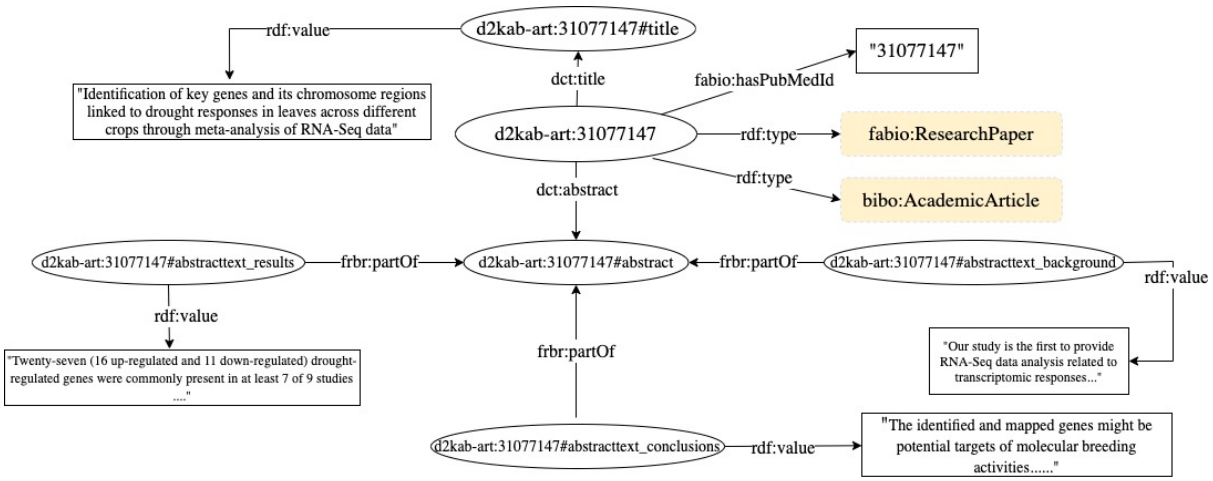


FIGURE 2 – Exemple de graphe RDF représentant une publication du corpus

lien *partie de* entre un résumé et une publication et aussi entre un résumé et ses sous-sections. La figure 2 illustre un graphe RDF représentant une publication scientifique avec un sous-ensemble de ses méta-données descriptives, à savoir le titre de la publication, le résumé et ses sous-sections (*background*, *results*, *conclusions*).

4.2 Description des annotations extraites des articles scientifiques

Une annotation A_i est une indication qu'une mention m_e d'une entité nommée e a été identifiée dans le résumé (ou l'une de ses sous-sections) d'un article a à une position de début d et une position de fin f. Dans ce travail, nous réutilisons le vocabulaire OA pour représenter les annotations d'entités nommées. Ainsi, une annotation A_i est représentée comme une instance de la classe `oa:Annotation` et est décrite par les informations suivantes :

— A_i a une cible représentée comme l'objet de la propriété `oa:hasTarget`. Cette cible représente une occurrence d'une entité nommée e identifiée par la présence de sa mention m_e (*surface form*) entre une position de début d et une position de fin f dans le texte du résumé ou une de ses sous-sections.

— A_i a un corps `oa:hasBody` qui renvoie vers l'URI d'une entité e définie dans une ontologie ou un vocabulaire du domaine tels par exemple l'URI d'un concept WTO [11] ou d'une classe de taxons dans la taxonomie NCBI [6].

La figure 3 présente les différentes classes d'entités de notre graphe de connaissances. Comme le montre cette figure, ces classes sont inter-connectées par des relations sémantiques.

La figure 4 illustre un exemple d'annotation extraite à partir de la sous-section conclusion d'un résumé identifiée par l'URI `d2kab-art:31077147#abstracttext_conclusions`. Dans cette sous-section, la mention du trait '*drought resistance*' a été identifiée et mise en correspondance avec l'entité WTO `wto:WTO_0000311`. La position de début et de fin, ainsi que la valeur de la mention m_e sont indiquées respectivement par l'utilisation des propriétés : `oa:start`, `oa:end` et `oa:exact`. Cette représentation structurée en RDF permet de modéliser à la fois les portions des publications à partir desquelles une annotation a été extraite et d'autre part leur mise en correspondance avec des entités prédéfinies dans des ontologies et des vocabulaires du domaine. Ceci offrira la possibilitéakhaw aux chercheurs d'explorer les contextes d'occurrence et de co-occurrence des entités dans les textes scientifiques.

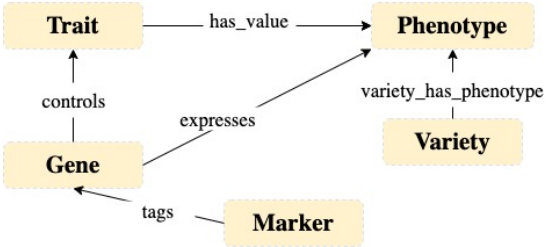


FIGURE 3 – Les classes d'entités extraites Alvis NLP et leurs relations sémantiques

5 Construction du graphe de connaissances

Nous présentons dans cette section le jeu de données et les outils utilisés pour la construction du graphe de connaissances. La figure 1 illustre le processus de construction graphe Wheat-KG à partir de la littérature scientifique de la génomique des plantes représentée par un corpus d'articles extraits à partir de PubMed [7].

5.1 Jeu de données

Le jeu de données fourni par l'équipe MaIAGE comprend plusieurs types d'informations stockées séparément dans

6. NCBI Taxonomy https://www.ncbi.nlm.nih.gov/taxonomy

7. https://pubmed.ncbi.nlm.nih.gov/

	Template d'URI
Entity	http ://ns.inria.fr/d2kab/{EntityClass}/{EntityID}
Article	http ://ns.inria.fr/d2kab/article/{PubmedId}
Annotation	http ://ns.inria.fr/d2kab/annotation/{annotationId}
Title	http ://ns.inria.fr/d2kab/article/{PubmedId}#title
Abstract	http ://ns.inria.fr/d2kab/article/{PubmedId}#abstract
Abstract section	http ://ns.inria.fr/d2kab/article/{PubmedId} #{sectionName}
Relation	http ://ns.inria.fr/d2kab/relation/{relationId}

TABLE 1 – Template de génération d'URI des ressources de notre graphe

Classe	
Nbre total d'annotations	88880
Nbre total d'articles	8496
Nbre total de gènes	1160
Nbre total de taxons	2462
Nbre total de phénotypes	98
Nbre total de marqueurs	521
Nbre total de variétés	77
Nbre total de relations	162

TABLE 2 – Nombre d'entités par classes dans notre graphe de connaissances

plusieurs fichiers CSV. Ainsi, ce jeu inclut un corpus constitué de 8496 publications à partir desquels un ensemble de 4318 entités nommées a été extrait en utilisant la plate-forme AlvisNLP qui offre une chaîne de traitement TAL pour l'annotation sémantique de documents textuels dans le domaine de la génomique des plantes. Cette plate-forme intégre plusieurs outils permettant la segmentation du texte en mots/phrases, la reconnaissance d'entités nommées, l'analyse de termes, le typage sémantique et l'extraction de relations présentes entre entités. Pour chaque publication du corpus, l'identifiant PubMed, le titre, le résumé et les possibles sous-sections du résumé sont fournis. Les entités nommées extraites par AlvisNLP ont été stockées dans un autre fichier CSV. Pour chaque occurrence d'entité, plusieurs informations sont renseignées sur plusieurs colonnes : l'identifiant Pubmed de l'article, la section du résumé où apparaît cette occurrence, la classe assignée à cette entité (gène, phénotype, marqueur, variété, taxon), la position (offset) de début et de fin de la mention de l'entité dans le texte. Enfin, les relations détectées par AlvisNLP sont stockées dans un troisième fichier CSV. Dans ce jeu, nous avons uniquement la relation *variety_has_phenotype* qui permet de relier une occurrence d'une entité nommée de type variété à une occurrence d'une entité nommée de type phénotype.

5.2 Processus de lifting

Pour transformer les annotations produites par AlvisNLP en un graphe RDF, nous avons utilisé l'outil morph-xR2rml [10]. Tout d'abord, nous avons défini un ensemble de règles de mapping. Ces règles décrivent un ensemble de *TripleMap* respectant le vocabulaire et la syntaxe qui sont fournis par la spécification xR2RML [8]. Chaque *TripleMap* définit un patron générique pour la génération de triplets RDF en respectant la modélisation proposée dans la section 4. L'ensemble des règles xR2RML définies sont disponibles dans le répertoire GitHub du projet [9]. La table 1 décrit les patrons pour génération d'URI des différentes ressources du futur graphe. De plus, dans l'objectif d'enrichir le graphe avec des méta-données descriptives des publications scientifiques, nous avons utilisé un micro-service SPARQL [10] qui permet d'interroger l'API PubMed Central et de récupérer les méta-données en RDF de chaque publication étant donné l'identifiant *PubMed* de la publication. Cette représentation RDF se base sur la modélisation présentée dans la section 4.1. Le tableau 2 représente le nombre de triplets pour chaque classe du graphe de connaissances.

6 Validation

Plusieurs CQ définies par un expert du domaine exprimant des besoins d'explorer la littérature scientifique selon différents critères sont présentées dans la section 3. Toutes les CQ ont été traduites en requêtes SPARQL [11] et les résultats ont été validés par l'expert. Cependant, nous nous contentons de présenter dans cette section uniquement les requêtes SPARQL implémentant les questions de compétence *CQ4* et *CQ5*. Ces requêtes montrent comment notre graphe de connaissances peut être exploité pour répondre aux besoins d'explorer la littérature discutant des gènes, des phénotypes et variétés de blé.

CQ4. Quelles sont les variétés de blé qui présentent un phénotype particulier ? L'intention de cette requête est de rechercher uniquement des variétés d'intérêt présentant le phénotype spécifié. En effet, dans le graphe de connaissances, les variétés sont inter-reliées à des phénotypes par la relation "`variety_has_phenotype`". le listing 1 présente la requête SPARQL correspondante à la CQ4.

```
SELECT distinct ?variety  ?document
WHERE {
?relation1 d2kab:hasVariety ?aVariety ;
            d2kab:hasPhenotype ?aPhenotype .
?aVariety a oa:Annotation ;
            oa:hasTarget ?t1 ;
            oa:hasBody ?Variety .
?t1 oa:hasSource ?partDoc1 .
?Variety a d2kab:Variety ;
            skos:prefLabel ?variety.
?aPhenotype a oa:Annotation ;
            oa:hasTarget ?t2 ;
            oa:hasBody ?Pheno .
?t2 oa:hasSource ?partDoc2 .
?Pheno skos:prefLabel 'drought tolerance' .
?partDoc1 frbr:partOf+ ?document .
?partDoc2 frbr:partOf+ ?document .
?document  a fabio:ResearchPaper .
}
```

Listing 1 – Requête SPARQL de la CQ4

8. https://www.i3s.unice.fr/~fmichel/xr2rml_specification_v5.html
9. https://github.com/Wimmics/d2kab-wheat-kg/tree/main/Mapping_rules
10. https://sparql-micro-services.org/service/pubmed/getArticleByPMId_sd/
11. https://github.com/Wimmics/d2kab-wheat-kg/tree/main/sparql-queries

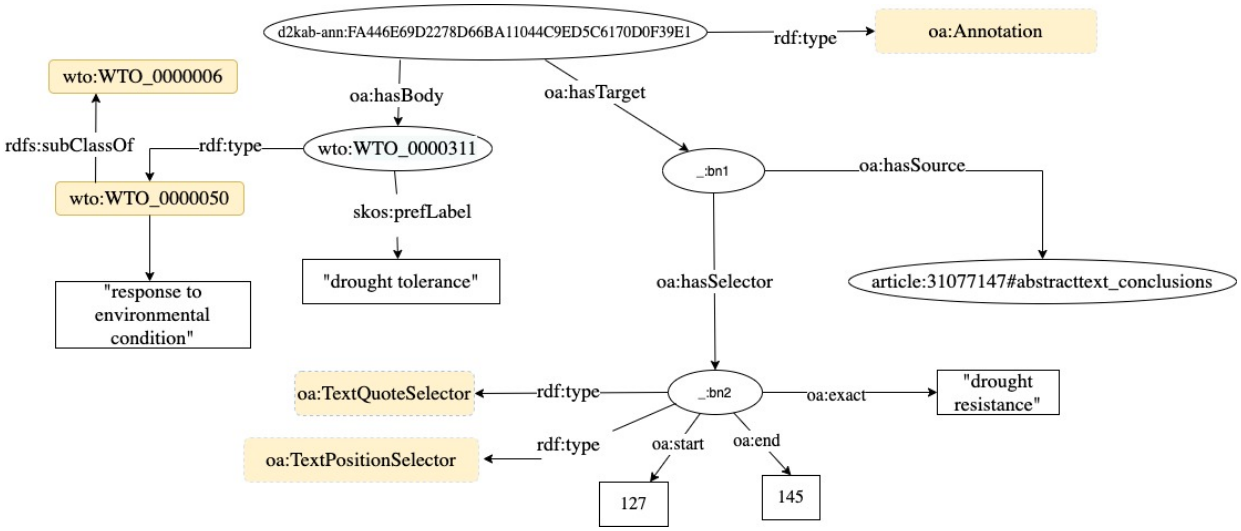


FIGURE 4 – Exemple de graphe RDF modélisant une annotation de l'entité *drought resistance* présente de la position 127 à 145 dans la sous-section conclusion du résumé d'un article

CQ5. Effectuer une recherche bibliographique de toutes les publications mentionnant des gènes spécifiques à des variétés de blé tendre (Triticum aestivum) qui présentent un phénotype général. Le résultat de cette requête devra inclure une liste de publications ayant mentionnées à la fois des gènes, une variété de blé tendre et un phénotype particulier. Si l'expert est intéressé par les gènes impliqués dans l'expression de phénotypes de résistance aux pathogènes, il faudrait alors inclure dans le résultat de la requête les phénotypes de résistance à tous les sous-types de pathogènes (bactérie, virus, champignons) qui peuvent obtenus a partir du thésaurus WTO. En effet, la requête du listing 2 exploite les relations hiérarchiques entre les concepts WTO pour inclure toutes les publications mentionnant des sous-concepts du phénotype '*pathogen resistance*'.

```
SELECT distinct  ?doc ?gene ?variety
WHERE {
?rel1 d2kab:hasVariety ?aVariety ;
    d2kab:hasPhenotype ?aPhenotype.

?aVariety a oa:Annotation ;
            oa:hasTarget ?t1 ;
            oa:hasBody ?Variety .
?t1 oa:hasSource ?d1 .
?Variety a d2kab:Variety ;
        skos:prefLabel ?variety.
?aPhenotype a oa:Annotation ;
            oa:hasTarget ?t2 ;
            oa:hasBody ?Phenotype .
?t2 oa:hasSource ?d2 .
?e2 skos:prefLabel "pathogen resistance" ;
    skos:narrower* ?Phenotype .
?aTaxon a oa:Annotation ;
        oa:hasTarget ?t ;
        oa:hasBody "Triticum aestivum".
?t oa:hasSource ?d .
?Taxon a d2kab:Taxon ;
        skos:prefLabel ?taxon.
?aGene a oa:Annotation ;
        oa:hasTarget ?t3;
        oa:hasBody ?Gene.
?t3 oa:hasSource ?d3 .
?Gene a d2kab:Gene ;
        skos:prefLabel ?gene.
?d1 frbr:partOf+ ?doc .
?d2 frbr:partOf+ ?doc .
?d3 frbr:partOf+ ?doc .
?d frbr:partOf+ ?doc .
?doc  a fabio:ResearchPaper .
}
```

Listing 2 – Requête SPARQL de la CQ5

7 Conclusion et Travaux futurs

Explorer la littérature scientifique en rapport avec les concepts clés la génomique des plantes peut s'avérer une tâche ardue pour les chercheurs. Ce travail de recherche s'attaque aux problèmes d'une recherche bibliographique transversale, et du liage des informations extraites à partir de la littérature scientifique avec les bases de données génomiques. En effet, la disponibilité de ressources sémantiques dans ce domaine (ontologies, thésaurus) peut s'avérer d'une grande utilité pour annoter les textes scientifiques et extraire les entités nommées. Dans ce papier, nous avons conçu et construit un graphe de connaissances Wheat-KG en considérant les annotations extraites à partir d'un corpus de publications scientifiques. Ces annotations sont produites par la plate-forme AlvisNLP et portent sur différentes entités nommées de différents types incluant des gènes, des phénotypes, des marqueurs génétiques, des variétés et des taxons en rapport avec la génomique du blé. Dans Wheat-KG, les contextes d'apparition des différentes entités sont décrits et représentés d'une manière structurée en se basant sur l'utilisation conjointe des vocabulaires standards du Web sémantique (i.e, [14]) et des ontologies du domaine en question (i.e., [11]). Ceci a permis de les interroger de manière uniforme avec le langage SPARQL et surtout d'exploiter les contextes d'apparition pour découvrir des associations implicites entre ces entités. Comme travaux futurs, nous en-

visageons d'intégrer dans le graphe de connaissances des observations collectées par des professionnels du domaine. Ces observations décrivent les stades de croissances des plantes, la fréquence d'attaque de maladies pour certaines variétés, les localisations géographiques des parcelles de culture, les paramètres météorologiques, etc. L'objectif serait de permettre le développement de modèles combinant des connaissances émanant de la littérature scientifique et des données d'observations.

8 Remerciements

Ce travail a été réalisé dans le cadre du projet "Des Données aux Connaissances en Agronomie et Biodiversité (D2KAB–www.d2kab.org) financé par l'Agence Nationale de la Recherche (ANR-18-CE23-0017)

Références

[1] Alberto Anguita, Miguel Garcia-Remesal, Diana de la Iglesia, and Víctor Maojo. NCBI2RDF : enabling full RDF-based access to NCBI databases. *BioMed research international*, 2013 :983805, 01 2013.

[2] Mouhamadou Ba and Robert Bossy. Interoperability of corpus processing work-flow engines : the case of alvisnlp/ml in openminted. In *Proceedings of the Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability (INTEROP 2016) at LREC 2016*, pages 15–18, Portorož, Slovenia, 2016.

[3] Robert Bossy, Louise Deleger, Estelle Chaix, Mouhamadou Ba, and Claire Nédellec. Bacteria Biotope at BioNLP Open Shared Tasks 2019. In *5th Workshop on BioNLP Open Shared Tasks BioNLP-OST@EMNLP-IJCNLP 2019, Association for Computational Linguistics*, page np, Hong-Kong, Hong Kong SAR China, November 2019. Jin-Dong Kim and Claire Nédellec and Robert Bossy and Louise Deléger.

[4] Ian Davis and Richard Newman. Expression of core FRBR concepts in RDF. https://vocab.org/frbr/core.

[5] John M Giorgi and Gary D Bader. Towards reliable named entity recognition in the biomedical domain. *Bioinformatics*, 36(1) :280–286, 06 2019.

[6] Jin-Dong Kim, Jung-Jae Kim, Xu Han, and Dietrich Rebholz-Schuhman. Extending the evaluation of genia event task toward knowledge base construction and comparison to gene regulation ontology task. *BMC bioinformatics*, 16 :S3, 07 2015.

[7] Jin-Dong Kima, Karin Verspoorb, Michel Dumontierc, and K Bretonnel Cohend. Semantic representation of annotation involving texts and linked data resources. *Semantic Web journal*, 2015.

[8] Robert Leaman, Ritu Khare, and Zhiyong Lu. Challenges in clinical natural language processing for automated disorder normalization. *Journal of biomedical informatics*, 57 :28–37, 2015.

[9] José-Lázaro Martínez-Rodríguez, Aidan Hogan, and I. Lopez-Arevalo. Information extraction meets the semantic web : A survey. *Semantic Web*, 11 :255–335, 2020.

[10] Franck Michel, Loïc Djimenou, Catherine Faron-Zucker, and Johan Montagnat. Translation of relational and non-relational databases into RDF with xR2RML. In Valérie Monfort, Karl-Heinz Krempels, Tim A. Majchrzak, and Ziga Turk, editors, *WEBIST 2015 - Proceedings of the 11th International Conference on Web Information Systems and Technologies, Lisbon, Portugal, 20-22 May, 2015*, pages 443–454. SciTePress, 2015.

[11] Claire Nédellec, Liliana Ibanescu, Robert Bossy, and Pierre Sourdille. WTO, an ontology for wheat traits and phenotypes in scientific publications. *Genomics & Informatics*, 18, 2020.

[12] Nadeesha Perera, Matthias Dehmer, and Frank Emmert-Streib. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in Cell and Developmental Biology*, 8, 2020.

[13] Silvio Peroni and David Shotton. FaBiO and CiTO : Ontologies for describing bibliographic resources and citations. *Journal of Web Semantics*, 17 :33–43, 2012.

[14] Robert Sanderson, Paolo Ciccarese, and Benjamin Young. Web annotation ontology. https://www.w3.org/TR/annotation-vocab/, 2017.